IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Christian Peterhans
1 June 2024

Outline

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

Executive Summary

- For this report, data was collected from SpaceX and Wikipedia. After data wrangling, exploratory data analysis was performed. Maps were created for geospatial analysis and a dashboard war created for interactive analysis. Finally, machine learning models were compared in order to give the best prediction results of landing success for the first stage of SpaceX rockets after launch.
- EDA has shown that the launch site has a high influence on the landing success, suggesting that the sites CCAFS SLC-40 and KSC LC-39A are most promising. The analysis also shows, that higher payload mass, that are often requested by launches to higher orbits, lead to more successful landings.
- The comparison of machine learning models show that support vector machines and decision trees are best performing to predicts the landing outcome.

Introduction

- SpaceX launches are much less expensive compared to other providers. A launch of a Falcon 9 rocket costs 62 million dollar. Launches of other providers usually cost more than 165 million. The reason for SpaceX to more economical is that the first stage can be reuse, if it successfully lands after the launch.
- Since the launch costs heavily depend on the success of the first stage's landing, the following report shall answer the question, what key factors of a launch lead to a successful landing of the first stage.

Section 1

Methodology

Methodology

Executive Summary

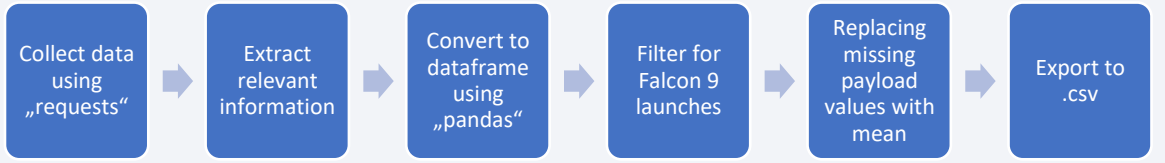
- Data collection methodology:
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

- Data was collected from two different sources with two different techniques
  - SpaceX API: https://api.spacexdata.com/v4/launches/past
  - Web scraping from Wikipedia: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Data Collection – SpaceX API

- URL to Notebook on GitHub: https://github.com/christianpeterhans/Applied-Data-Science-Capstone/blob/e2316566e8824b48a4f7364d3e5b70eca0492adc/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- URL to Notebook on GitHub: https://github.com/christianpeterhans/Applied-Data-Science-Capstone/blob/e2316566e8824b48a4f7364d3e5b70eca0492adc/jupyter-labs-webscraping.ipynb

Data Wrangling

- Replacement of missing payload values with mean
- Performing Exploratory Data Analysis (EDA)
  - Analyzing number of launch per site
  - Analyzing number and occurrence of each obit
  - Analyzing number and occurrence of mission outcome
- Determination of training labels
  - Creating landing outcome label

- URL to Notebook on GitHub: https://github.com/christianpeterhans/Applied-Data-Science-Capstone/blob/e2316566e8824b48a4f7364d3e5b70eca0492adc/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
- Charts that were used for analysis
  - Scatter plot with **flight number** and **payload mass** on the axes, the color indicates if the landing was successful → the plot shows the relationship between flight number, payload mass and success
  - Scatter plot with **flight number** and **launch site** on the axes, the color indicates if the landing was successful → the plot shows the relationship between flight number, launch site and success
  - Scatter plot with **payload mass** and **launch site** on the axes, the color indicates if the landing was successful → the plot shows the relationship between payload mass, launch site and success
  - Bar chart with **success rate** over **orbit** → the plot shows launches to which orbit were most successful
  - Scatter plot with **flight number** and **orbit type** on the axes, the color indicates if the landing was successful → the plot shows the relationship between flight number, orbit type and success
  - Scatter plot with **payload mass** and **orbit type** on the axes, the color indicates if the landing was successful → the plot shows the relationship between payload mass, orbit type and success
  - Line plot with success rate over time (year) → the plot shows the evolution of successful landing over time
- URL to Notebook on GitHub:
  https://github.com/christianpeterhans/Applied-Data-Science-Capstone/blob/e2316566e8824b48a4f7364d3e5b70eca0492adc/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- SQL queries performed on data
  - Unique launch sites
  - 5 records, where launch site begins with “CCA”
  - Total payload mass carried by boosters launched by NASA
  - Average payload mass carried by booster version F9 v1.1
  - Date of first successful landing outcome in ground pad
  - Names of boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Total number of successful and failure mission outcomes
  - Names of booster version which have carried the maximum payload mass
  - Month names, landing outcomes in droneship, booster version, lauch site for the months in 2015
  - Count of landing outcomes between 2010-06-04 and 2017-03-20
- URL to Notebook on GitHub:
  https://github.com/christianpeterhans/Applied-Data-Science-Capstone/blob/e2316566e8824b48a4f7364d3e5b70eca0492adc/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

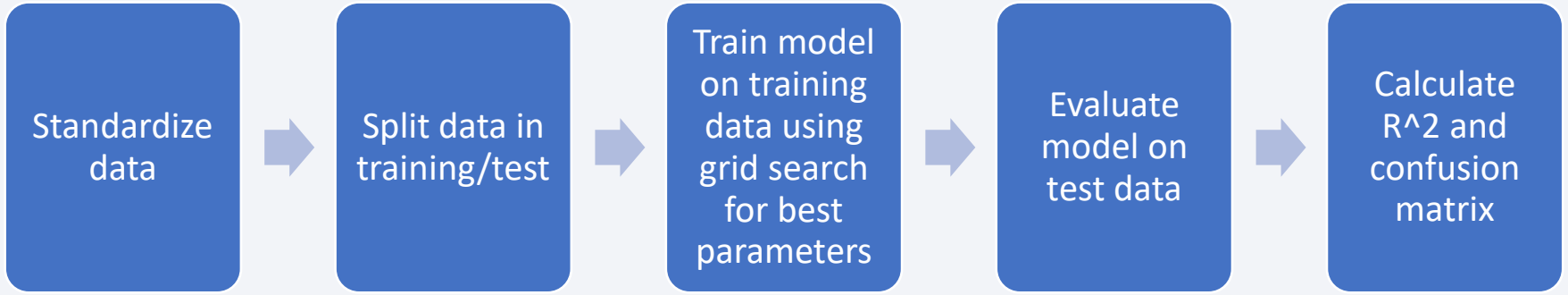
- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
- Added a circles and markers at each launch site → to see where the launch sites are
- Colored the marker with successful landing green and with unsuccessful landing red → to see how many successful landing a launch site has
- Created lines indicating distance to sea → to find out if distance to sea influences success
- Explain why you added those objects
- URL to Notebook on GitHub: https://github.com/christianpeterhans/Applied-Data-Science-Capstone/blob/e2316566e8824b48a4f7364d3e5b70eca0492adc/lab_jupyter_launch_site_location.ipynb

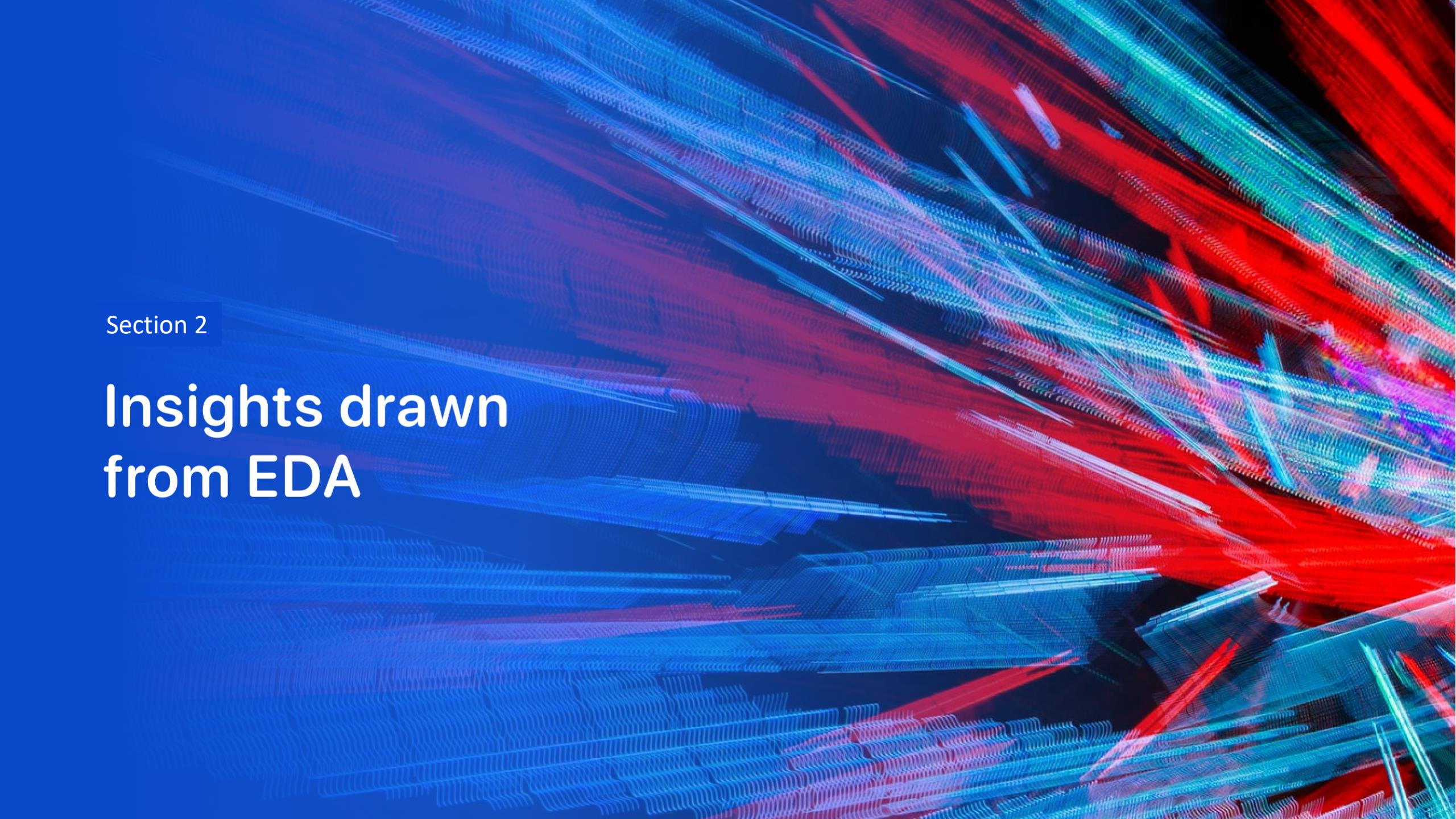
Build a Dashboard with Plotly Dash

- Dashboard contains the following plot and interactions
  - Drop down to choose launch site
  - Pie chart with success rate
  - Slider with payload mass
  - Scatter plot with payload mass and landing success on axes and the color indicates the booster version
  - These elements were used to analysis the success rates of the different launch sites and if the success rate at a launch site is influenced by the payload and by the booster version
- URL to Notebook on GitHub:
  https://github.com/christianpeterhans/Applied-Data-Science-Capstone/blob/2a2fd5a6e39aef58ee33e4e2fe86b8b3dbaad957/Build_a_Dashboard_Application_with_Plotly_Dash.py

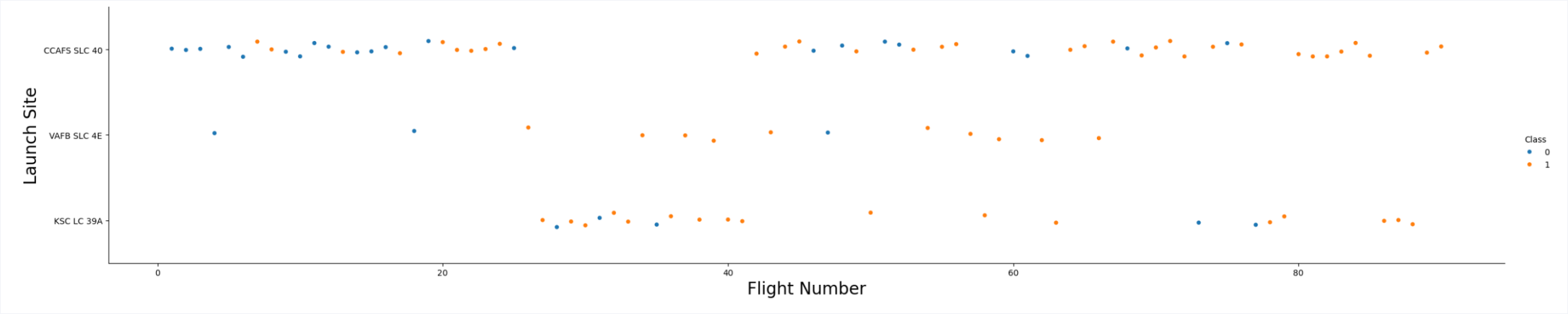
Predictive Analysis (Classification)

- Standardized features using sklearn
- Split data in training (80%) and test data (20%)
- Compared following models to predict landing success on test data. Comparison based on R^2 values and confusion matrix
  - Logistic regression
  - Support vector machine
  - Decision tree classifier
  - K nearest neighbor
- URL to Notebook on GitHub: https://github.com/christianpeterhans/Applied-Data-Science-Capstone/blob/2a2fd5a6e39aef58ee33e4e2fe86b8b3dbaad957/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



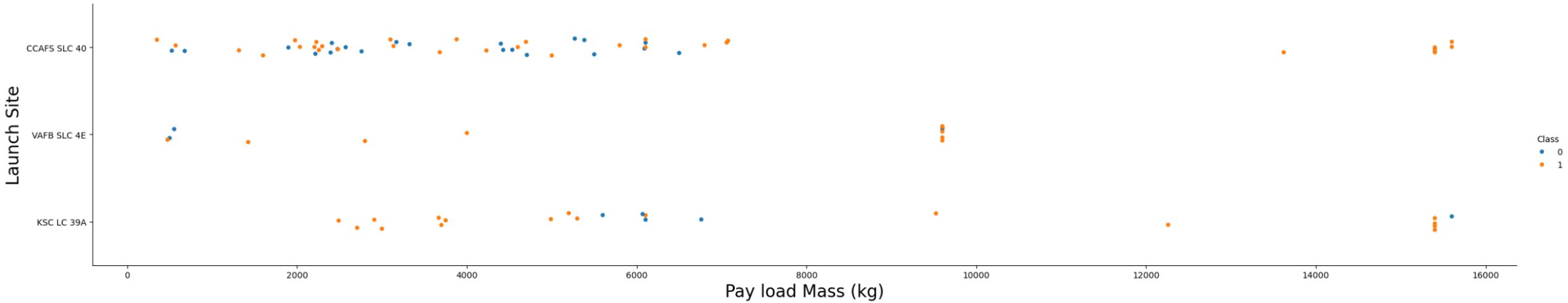

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



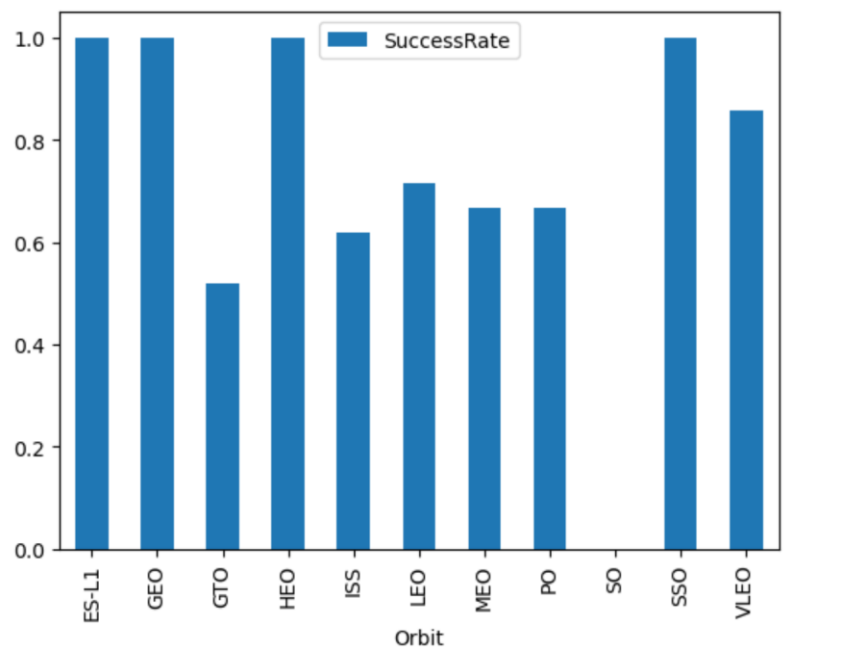
- With higher flight number, more successful landings occured
- VAFB SLC 4E and KSC LC 39A have higher success rates but less launches than CCAFS SLC 40
- VAFB SLC 4E wasn‘t used for most recent launches

Payload vs. Launch Site



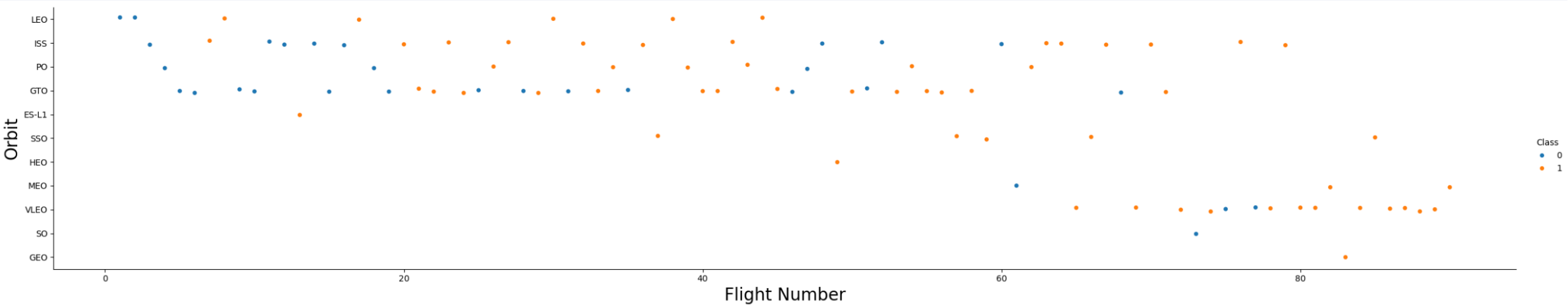
- The launches with the highest payloads were not launched from VAFB SLC 4E
- The landings of launches with high payloads are very successful
- The successrate with low payload is worst at CCAFS SLC 40

Success Rate vs. Orbit Type



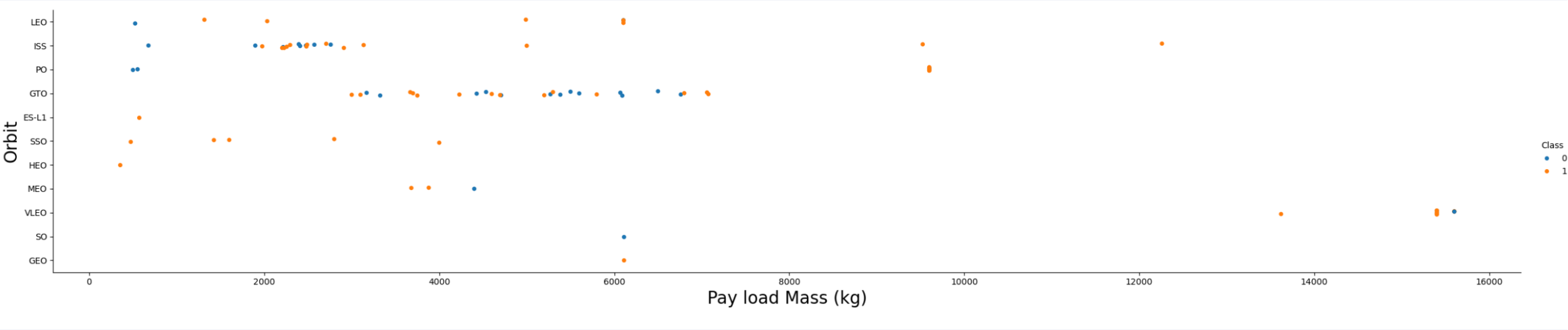
- The highest success rates have launches to the orbits : ES-L1, GEO, HEO and SSO
- The lowest success rates have launches to the orbit SO

Flight Number vs. Orbit Type



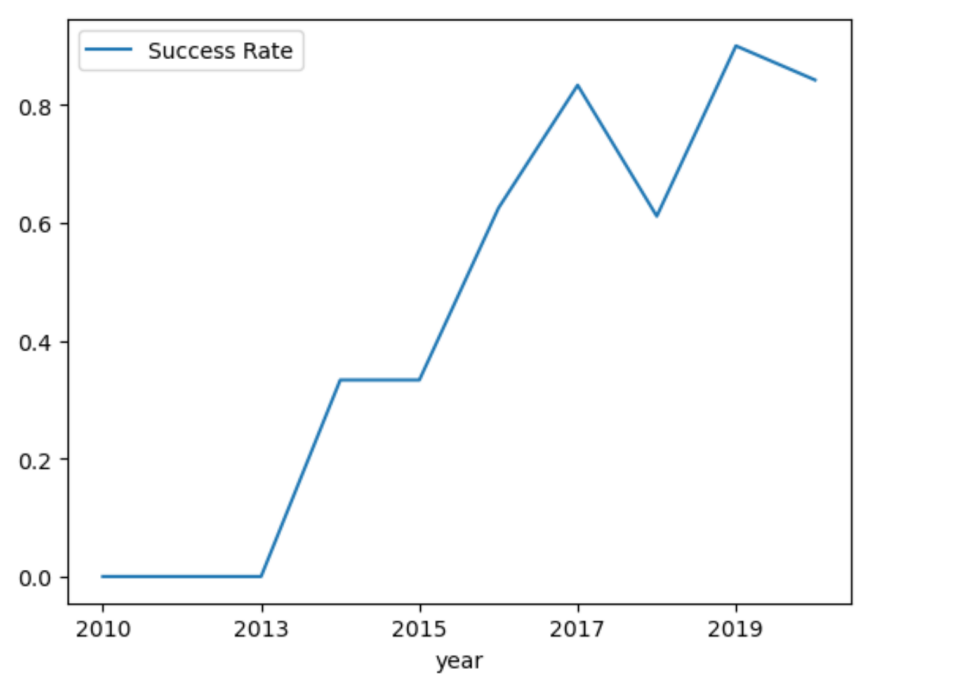
- The extreme success rates occur, where very little launches were performed
- ISS and GTO have a high number of launches

Payload vs. Orbit Type



- To VLEO orbit, the highest payloads were transported
- The payloads to GTO were higher than to ISS
- High payloads have high success rates

Launch Success Yearly Trend



- The success rate massively increased over the years
- It started around 0 and ended at around 0.8

All Launch Site Names

- These are all the launch sites, that are present in the data

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

- 4 different launch sites in data

Launch Site Names Begin with 'CCA'

- 5 records, where launch site name begins with CCA

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- The first 5 records are the exact same launch site

Total Payload Mass

- The total payload mass launched by NASA is

SUM(PAYLOAD_MASS__KG_)
99980

Average Payload Mass by F9 v1.1

- The average payload mass for a F9 v1.1 booster is

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

- Roghly 2.5 tons

First Successful Ground Landing Date

- Date of the first successful landing outcome on ground pad

Date
2015-12-22

- Already 2015

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

Booster_Version
F9 FT B1029.1
F9 FT B1036.1
F9 B4 B1041.1

- 3 different boosters

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcome

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

- Only 1 failure compared to 100 successful missions

Boosters Carried Maximum Payload

- Names of the booster which have carried the maximum payload mass

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

- 12 different boosters

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

- 2 failed landings in drone ship in 2015 with different boosters

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

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

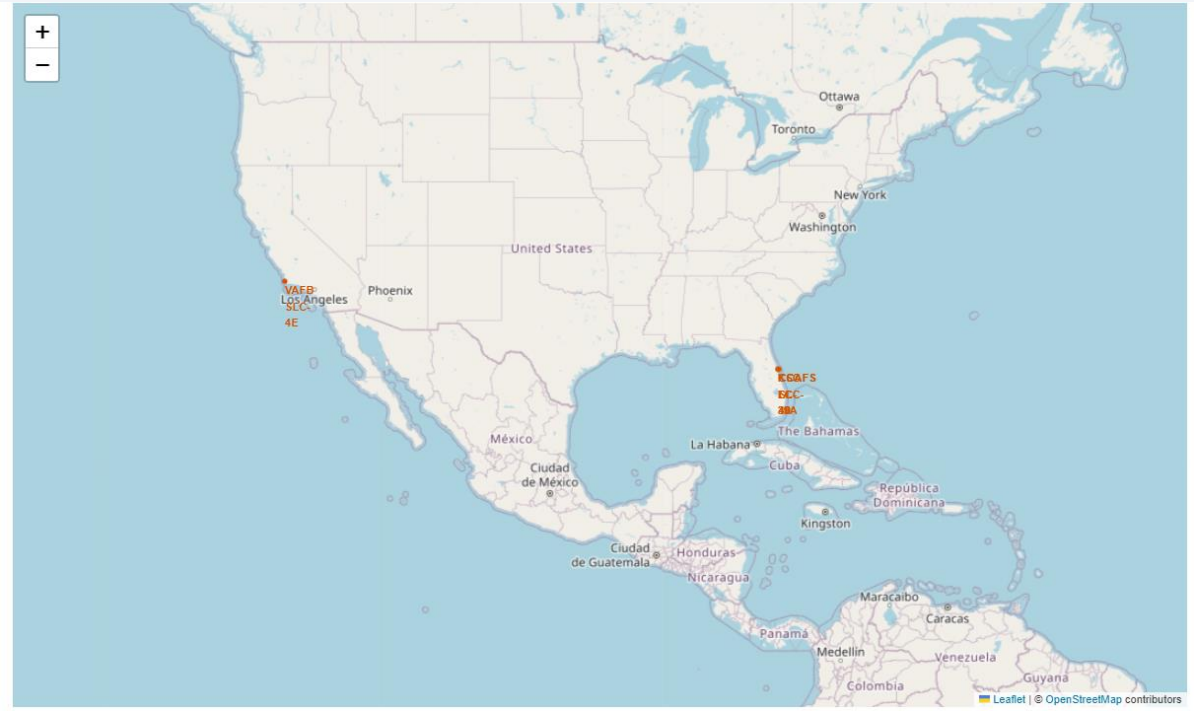
Landing_Outcome	COUNT(Landing_Outcome)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

- Most often no attempt

Section 3

Launch Sites Proximities Analysis

Launch site overview



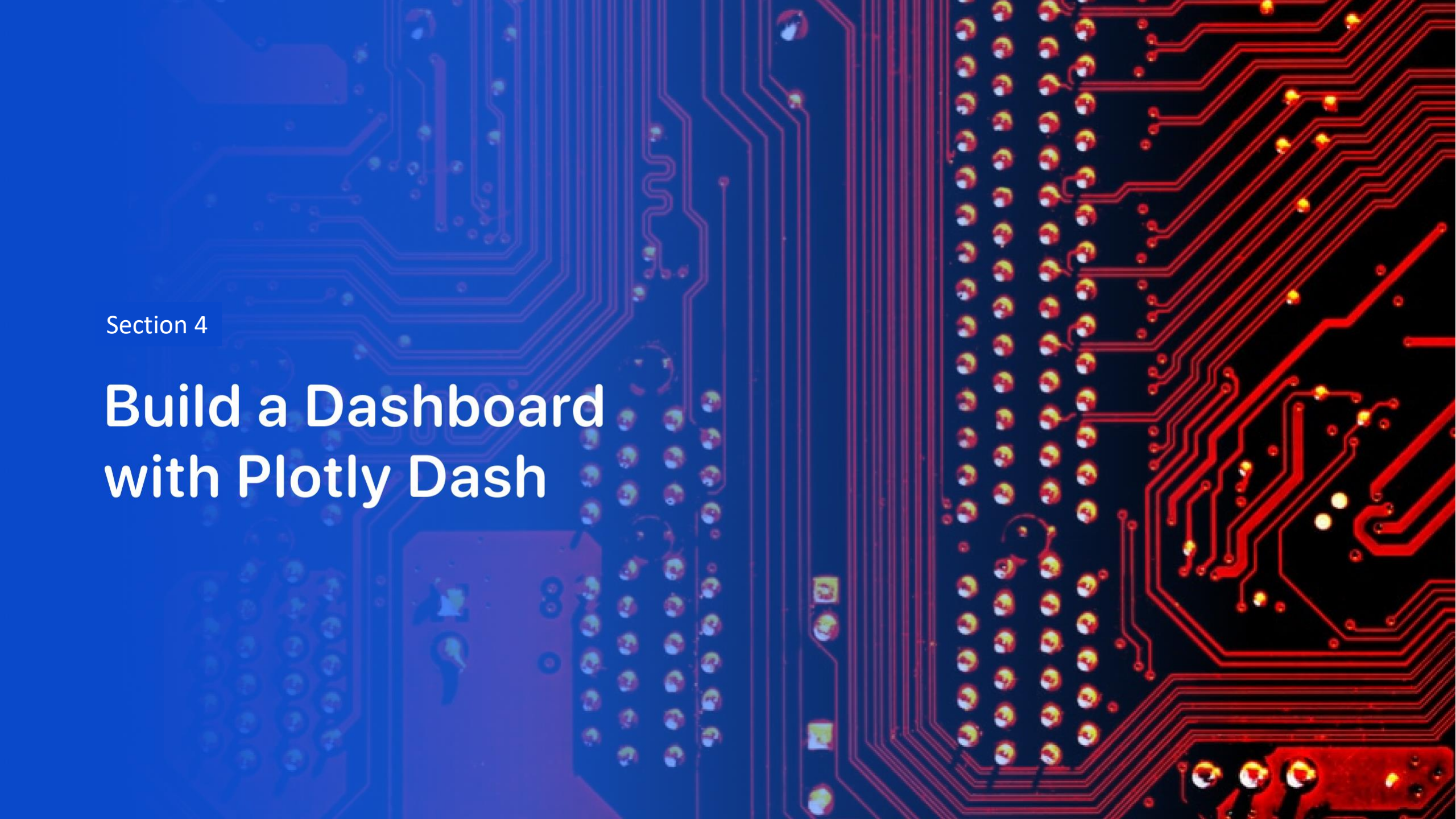
- The launch sites are either in the east or the west of the USA
- All launch sites are close to an ocean

KSC LC-39A

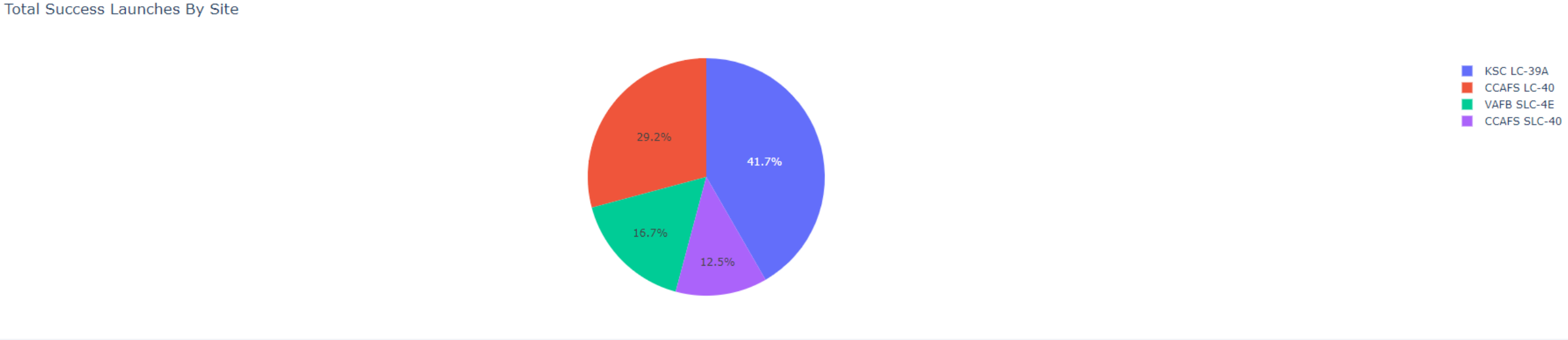
- KSC LC-A39 has the highest success rate of around 77%

Distance from CCAFS SLC-40 to coastline

- The distance between CCAFS SLC-40 and the coastline is only 0.95km

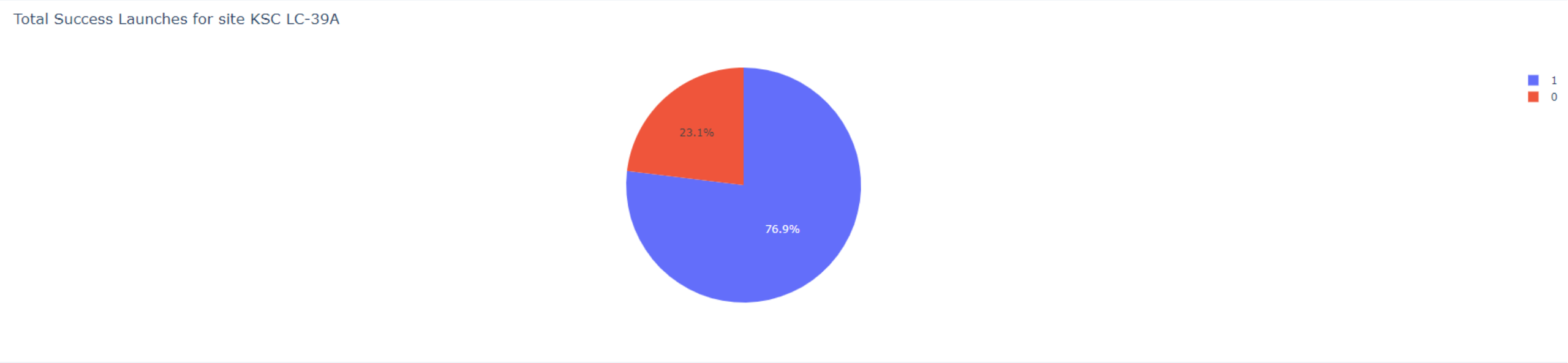

Section 4
Build a Dashboard with Plotly Dash

Total successful landings by site



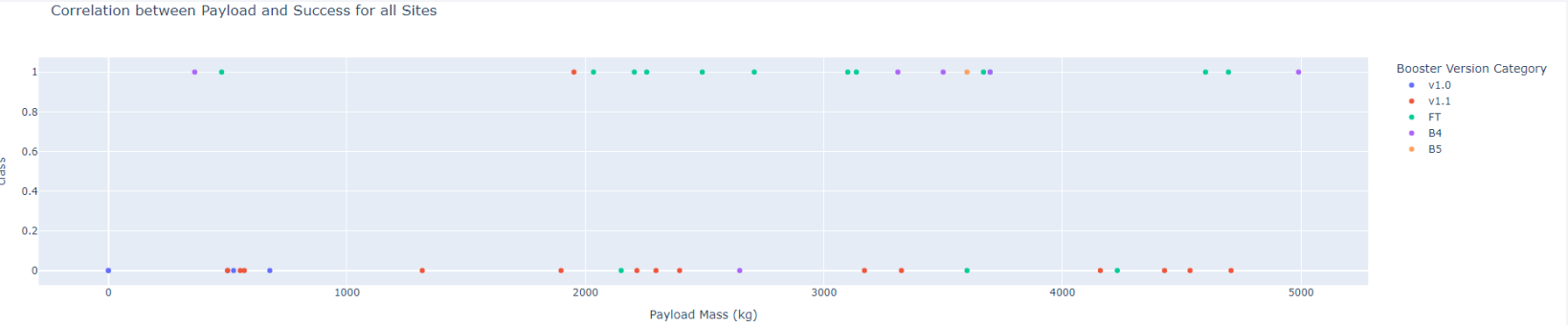
- KSC LC39A has the highest number of successful landings
- CCAFS SLC-40 has the least successful landings

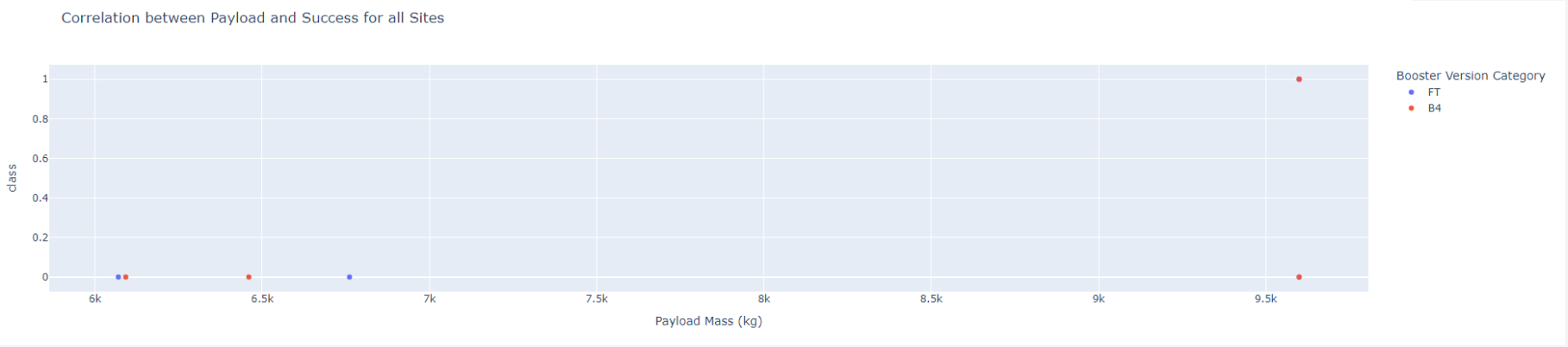
Launches from KSC LC-39A



- The highest success rate has the launch site KSC LC-39A

Payload vs Success



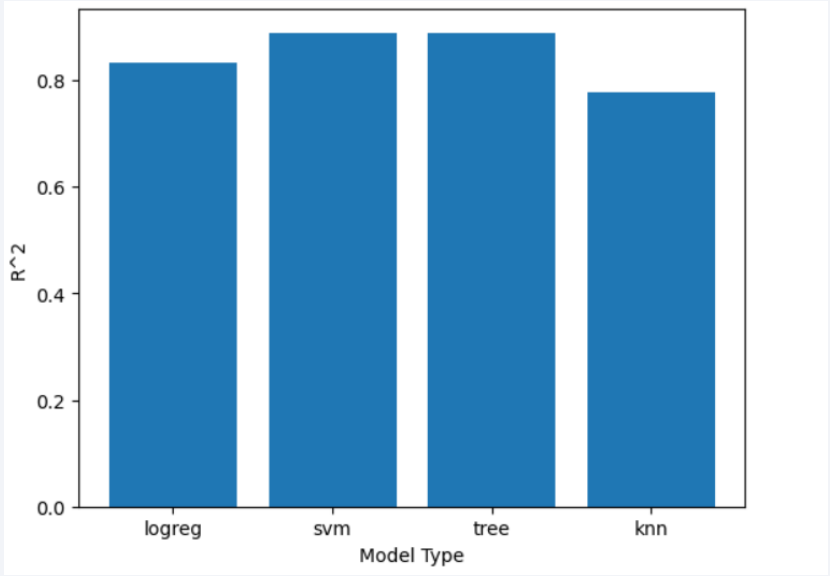


- For all ranges of payload, successful and failed landings ocur

Section 5

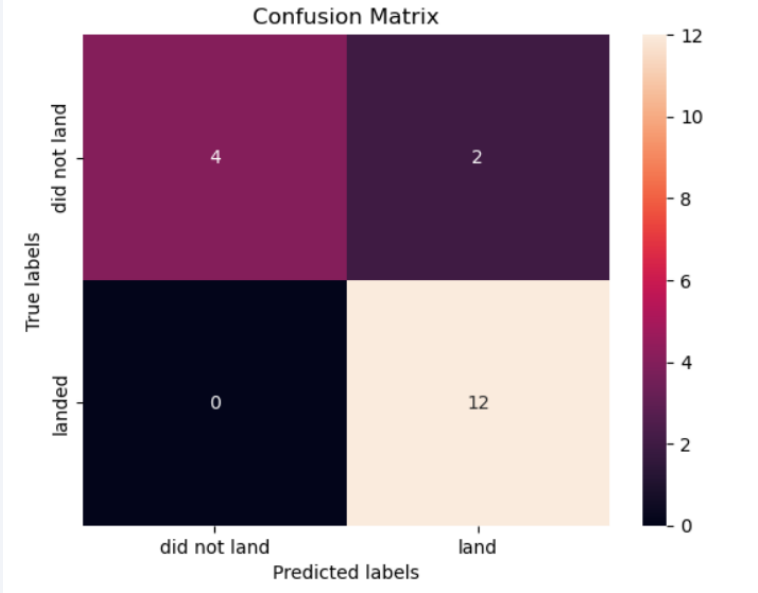
Predictive Analysis (Classification)

Classification Accuracy



- Support vector machines and decision trees perform equally good on the test data

Confusion Matrix



- Interestingly, svm and decision tree also share the same confusion matrix

Conclusions

- Launches with higher flight number had more success. For the most recent and most successful launches, only the launching sites CCAFS SLC-40 and KSC LC-39A were used, indicating that SpaceX has judged the situation similarly.
- High payload masses at the launch lead to a higher success rate for landing
- Launches to a higher orbit tend to require higher payload masses and thus lead to a higher success rate for landing
- KSC LC-39A is the launch pad with the highest success rate for landing

Appendix

- Machine Learning code
  - Standardization

```
]: # students get this
transform = preprocessing.StandardScaler()
X = transform.fit_transform(X)
```

  - Train-Test split

```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=1)
```

  - Grid search for logistic regression

```
parameters ={'C':[0.01,0.1,1],
             'penalty':['l2'],
             'solver':['lbfgs']}
```

```
parameters ={"C":[0.01,0.1,1],'penalty':['l2'], 'solver':['lbfgs']}# l1 lasso l2 ridge
lr=LogisticRegression()
logreg_cv = GridSearchCV(lr,
                         param_grid=parameters,
                        cv=10)
logreg_cv.fit(X_train,Y_train)
```

Thank you!