IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sebastian Korab
20th December 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data was collected based on SpaceX REST API and web scraping
  - Data was filtered, missing values were managed and one-hot-encoding was applied
  - Exploratory data analysis (EDA) was performed using visualization and SQL
  - Interactive visual analytics was performed using Folium and Plotly Dash
  - After segregation in training and test sets, predictive models were built. Different methods and its results were compared to each other.
- Summary of all results
  - Mission success rate increased with the number of launches and over time.
  - Launches aiming at Orbits ES-L1, GEO, HEO and SSO had a success rate of 100 %.
  - The success rate is the highest for payloads between 2000kg and 5200kg. Launch site KSC LC-39A has the highest success rate for launches (76.9 %).
  - The decision tree is the best predictive model, but false positive results in the confusion matrix are a problem and need to be reduced/eliminated.

Introduction

Project background

Technological progress has led to growing demand in satellites. As government led programs come along with higher costs per rocket launch, SpaceX has defined the commercialized cost-efficient rocket launches as one of its goals.

SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Goal

This project aims at analyzing historical data from the SpaceX launches in order to identify drivers, such as launch site or targeted orbits, for positive outcomes (i.e. launches with successful landings).

Based on the exploratory data analysis we will build a model for the prediction of a positive landing.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected based on SpaceX REST API and web scraping
- Perform data wrangling
  - Data was filtered, missing values were managed and one-hot-encoding was applied
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - After segregation in training and test sets, predictive models were built. Different methods and its results were compared to each other.

Data Collection

Data was collected based on SpaceX REST API and web scraping

SpaceX REST API

- requesting and parsing the SpaceX launch data using the GET request
- decoding the response by the use of the .json()-function
- filtering the dataframe, so that only Falcon 9 launches will remain

Web Scraping

- requesting Falcon 9 launch data from Wikipedia
- creation of a `BeautifulSoup` object from the HTML `response`
- extracting all column/variable names from the HTML header
- creation of a dataframe by parsing the launch HTML tables

Data Collection – SpaceX API

- requesting and parsing the SpaceX launch data, the response got encoded and the dataframe got filtered I order to keep only Falcon 9 launches

- GitHub URL to the SpaceX API calls notebook:

https://github.com/YoSebastian/SebastiannKorab_SpaceY_IBM_DataScience_Capstone/blob/main/EDIT_jupyter-labs-spacex-data-collection-api.ipynb

```
# Takes the dataset and uses the cores column to call the API and append the data to the lists
def getCoreData(data):
    for core in data['cores']:
            if core['core'] != None:
                response = requests.get("https://api.spacexdata.com/v4/cores/"+core['core']).json()
                Block.append(response['block'])
                ReusedCount.append(response['reuse_count'])
                Serial.append(response['serial'])
            else:
                Block.append(None)
                ReusedCount.append(None)
                Serial.append(None)
            Outcome.append(str(core['landing_success'])+' '+str(core['landing_type']))
            Flights.append(core['flight'])
            GridFins.append(core['gridfins'])
            Reused.append(core['reused'])
            Legs.append(core['legs'])
            LandingPad.append(core['landpad'])
```

Data Collection - Scraping

- requesting Falcon 9 launch data, creation of a `BeautifulSoup` and a dataframe by parsing the launch HTML tables
- GitHub URL to the Web Scraping notebook:

https://github.com/YoSebastian/SebastianKorab_SpaceY_IBM_DataScience_Capstone/blob/main/EDIT_jupyter-labs-webscraping.ipynb

```
launch_dict= dict.fromkeys(column_names)

# Remove an irrelvant column
del launch_dict['Date and time ( )']

# Let's initial the launch_dict with each value to be an empty list
launch_dict['Flight No.'] = []
launch_dict['Launch site'] = []
launch_dict['Payload'] = []
launch_dict['Payload mass'] = []
launch_dict['Orbit'] = []
launch_dict['Customer'] = []
launch_dict['Launch outcome'] = []
# Added some new columns
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]
```

Data Wrangling

- exploratory data analysis was conducted and training labels were determined
- calculation of:
  - the number of launches on each site
  - the number and occurrence of each orbit
  - the number and occurrence of mission outcome of the orbits
- creating a landing outcome label

- GitHub URL to the data wrangling notebooks:

https://github.com/YoSebastian/SebastianKorab_SpaceY_IBM_DataScience_Capstone/blob/main/EDIT_labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Visualization of:
  - the relationship between Flight Number and Launch Site
  - the relationship between Payload and Launch Site
  - the relationship between success rate of each orbit type
  - the relationship between Flight Number and Orbit type
  - the relationship between Payload and Orbit type
  - the launch success yearly trend
- GitHub URL to the EDA with data visualization notebooks:

https://github.com/YoSebastian/SebastianKorab_SpaceY_IBM_DataScience_Capstone/blob/main/EDIT_jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- SQL queries leading to:
  - the names of the unique launch sites in the space mission
  - 5 records where launch sites begin with the string 'CCA'
  - 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
  - names of the booster versions which have carried the maximum payload mass
  - ranking the count of landing outcomes
- GitHub URL to the EDA with SQL notebook:

https://github.com/YoSebastian/SebastianKorab_SpaceY_IBM_DataScience_Capstone/blob/main/EDIT_jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Added to folium map:
  - circles for all launch sites on a map
  - colored markers the success(green)/failed(red) launches for each site on the map
  - lines for the distances between a launch site to its proximities
- The added items provide an overview of the relevant launch locations, success rates and resulting risks from failed launches affecting proximities, such as cities or highways
- GitHub URL to the Folium interactive map notebook:

https://github.com/YoSebastian/SebastianKorab_SpaceY_IBM_DataScience_Capstone/blob/main/EDIT_lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Following plots/graphs were added to a dashboard:
  - dropdown list to enable Launch Site selection
  - pie chart to show the total successful launches count for all sites
  - slider to select payload range
  - scatter chart to show the correlation between payload and launch success
- GitHub URL to the Plotly Dash Python file:

https://github.com/YoSebastian/SebastianKorab_SpaceY_IBM_DataScience_Capstone/blob/main/spacex_dash_app.py

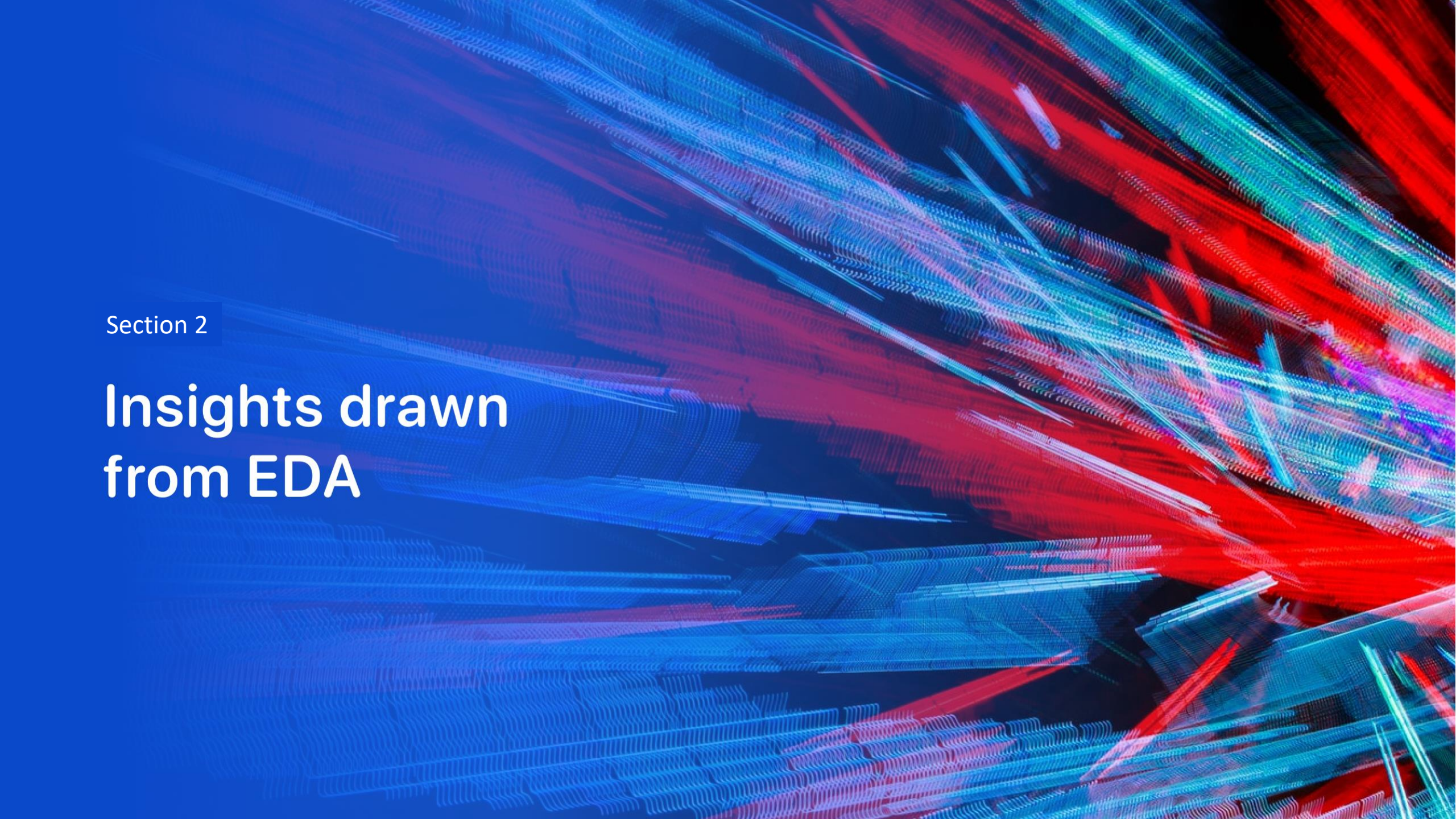
Predictive Analysis (Classification)

- performed EDA and determined training labels
- created a column for the class (successful mission/failure)
- standardized the data
- split into training data and test data
- found the best hyperparameters for SVM, Classification Trees and Logistic Regression
- found the method performs best using test data
- GitHub URL to the Predictive Analysis Python file:

https://github.com/YoSebastian/SebastianKorab_SpaceY_IBM_DataScience_Capstone/blob/main/EDIT_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

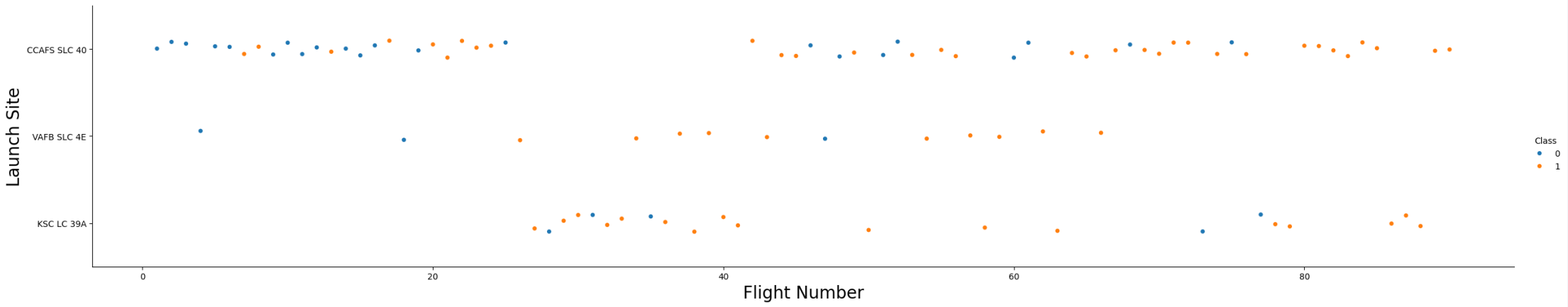
Results

- Exploratory data analysis results
  - Rate for successful missions has increased over time
  - Landing site KSC LC39A is the safest landing site
  - Launches aiming at Orbits ES-L1, GEO, HEO and SSO were all successful
- Interactive analytics demo in screenshots
  - Launch locations are chosen, which mitigate the risk from failed launches affecting proximities, such as cities or highways. All launch locations are close to the sea (Pacific Ocean, Atlantic Ocean) and rather in the south of the United States.
- Predictive analysis results
  - The decision tree model was identified as the model for the predictive analysis


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

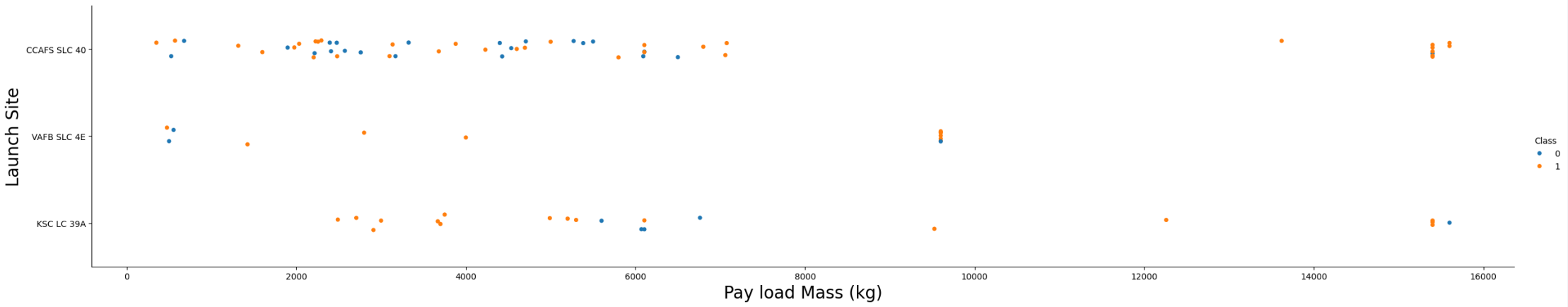
Scatter plot of Flight Number vs. Launch Site



Generally, the success rate increases with increasing Flight number. It seems that the Launch Site VAFB SLC 4E got decommissioned after Flight Number 70. CCAFS SLC 40 is used more as a Launch Site compared to KSC LC 39A.

Payload vs. Launch Site

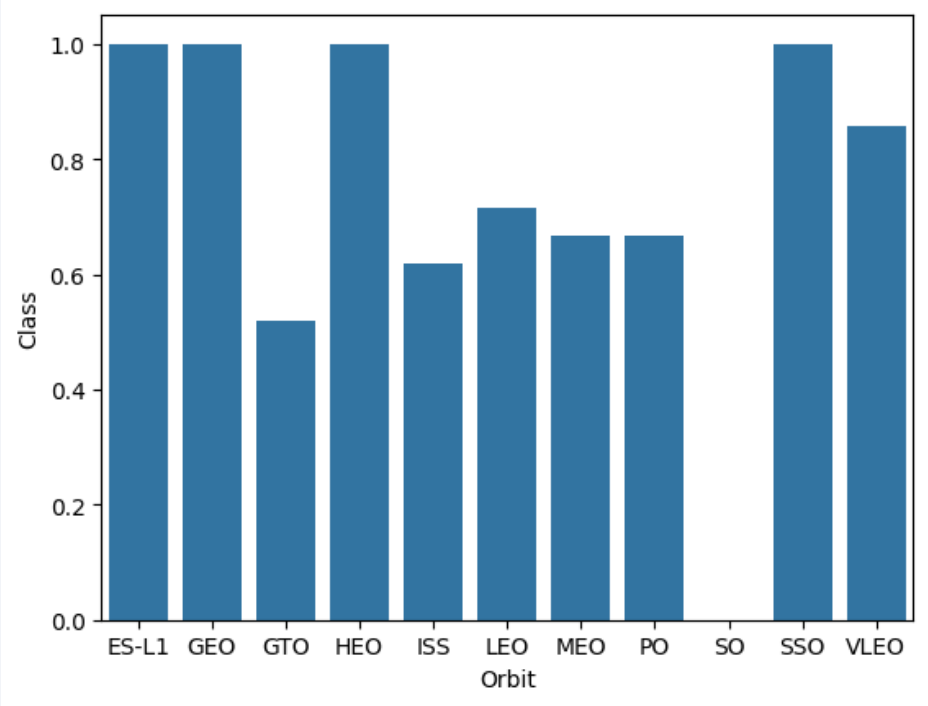
Scatter plot of Payload vs. Launch Site



Looking at the Payload Vs. Launch Site scatter point chart you will find that there are no rockets launched for heavy payload mass (greater than 10000) from launch site VAFB-SLC.

Success Rate vs. Orbit Type

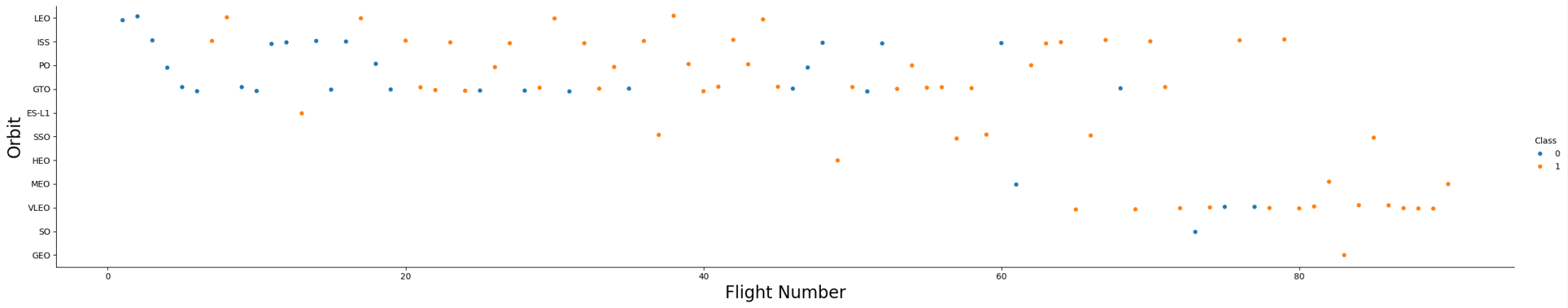
Bar chart for the success rate of each orbit type



Launches aiming at Orbits ES-L1, GEO, HEO and SSO had a success rate of 100 %

Flight Number vs. Orbit Type

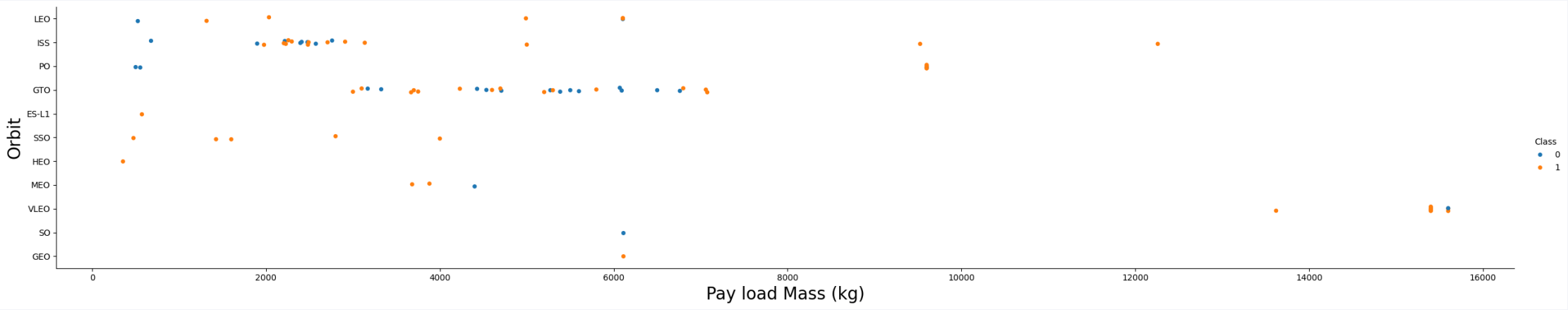
Scatter plot of Flight number vs. Orbit type



Success for launches aiming at the LEO orbit appears to be related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type

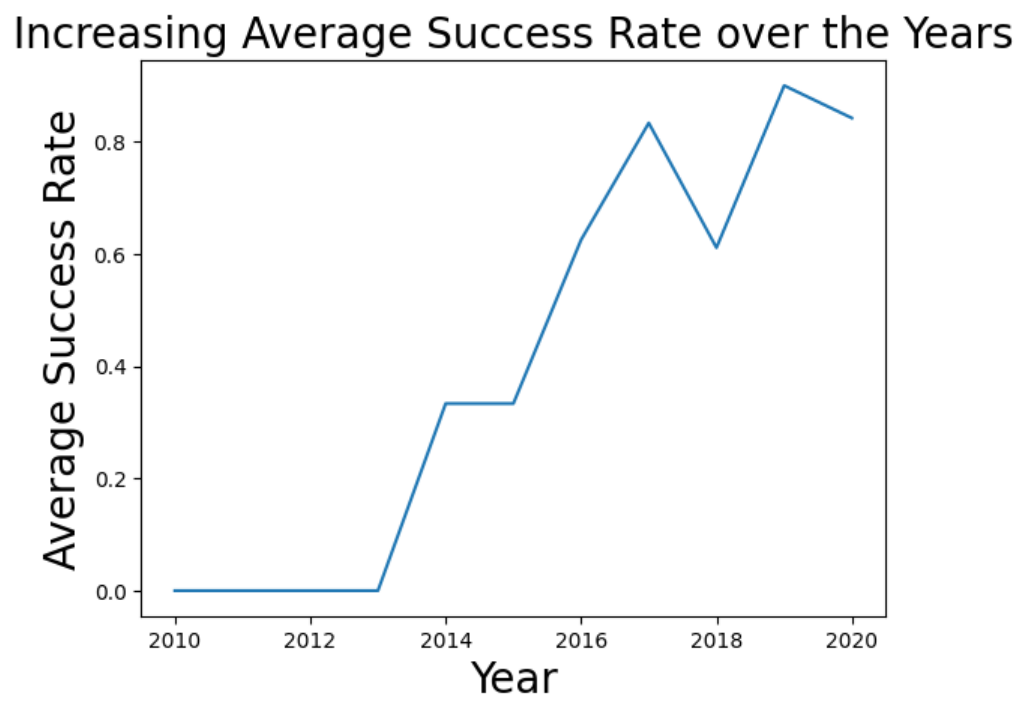
Scatter plot of Payload vs. Orbit type



The rate for successful landings increases with heavier payloads with launches aiming at the orbit types POLAR, LEO and ISS.

Launch Success Yearly Trend

Lines chart of yearly average success rate



The rate for successful mission increased over the years. It stayed stable in 2014 and had a slight drop in 2018, but a positive trend can be seen.

All Launch Site Names

- Identified launch site names are:

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

- %sql SELECT Distinct(LAUNCH_SITE) FROM SPACEXTBL;

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`:

Booster_Version	Launch_Site	Payload
F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit
F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese
F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2
F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1
F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2

- %sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE'CCA%' LIMIT 5;

Total Payload Mass

- Calculated total payload mass (in kg) carried by boosters from NASA:

SUM(PAYLOAD_MASS__KG_)
45596

- %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER='NASA (CRS)';

Average Payload Mass by F9 v1.1

- Calculated average payload mass (in kg) carried by booster version F9 v1.1:

AVG(PAYLOAD_MASS__KG_)
2928.4

- %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version='F9 v1.1';

First Successful Ground Landing Date

- Date of first successful landing outcome on ground pad:

MIN(DATE)
2015-12-22

- %sql SELECT MIN(DATE) FROM SPACEXTBL WHERE Landing_Outcome='Success (ground pad)';

Successful Drone Ship Landing with Payload between 4000 and 6000

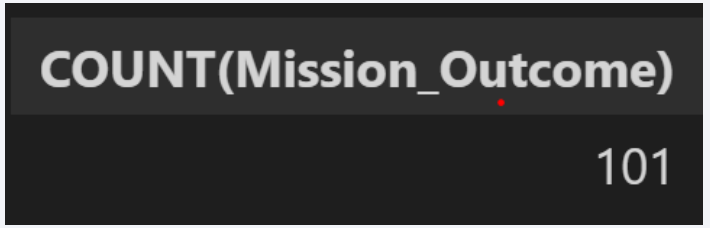
- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- %sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE Landing_Outcome='Success (drone ship)' and PAYLOAD_MASS__KG_>=4000 and PAYLOAD_MASS__KG_<=6000 ;

Total Number of Successful and Failure Mission Outcomes

- Calculated total number of successful and failed mission outcomes:



- %sql SELECT COUNT(Mission_Outcome) FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE 'Success%' OR MISSION_OUTCOME LIKE 'Failure%'

Boosters Carried Maximum Payload

- Names of boosters which have carried the maximum payload mass:
  - 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
- %sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);

2015 Launch Records

- List of failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015:

Booster_Version	Launch_Site	Landing_Outcome	MONTH
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)	01
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)	04

- %sql SELECT BOOSTER_VERSION, LAUNCH_SITE, LANDING_OUTCOME, substr(DATE,6,2) as MONTH FROM SPACEXTBL WHERE Landing_Outcome='Failure (drone ship)' and substr(DATE,0,5)='2015';

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

- 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:

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

- %sql SELECT COUNT(LANDING_OUTCOME) \

  as COUNTS, LANDING_OUTCOME \

  FROM SPACEXTBL \

  WHERE DATE between '2010-06-04' and '2017-03-20' \

  group by LANDING_OUTCOME \

  order by COUNTS DESC;

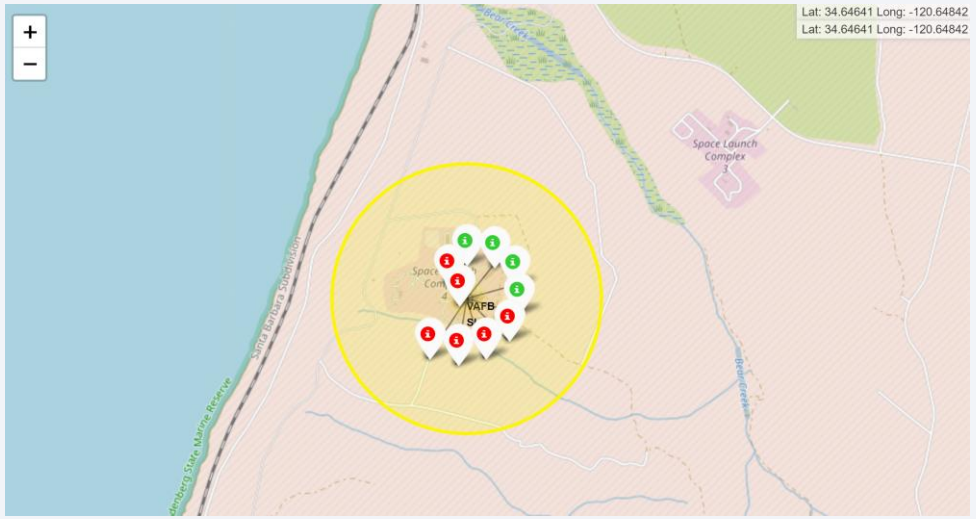

Section 3
Launch Sites
Proximities Analysis

Folium Map – Launch Sites

- All launch locations are close to the sea (Pacific Ocean, Atlantic Ocean) and rather in the south of the United States – the reason is that southern locations are closer to the equator and the earth's rotation can be used to reduce fuel consumption during the launch.

- There are 46 unique launch coordinates in Florida and 10 in California.

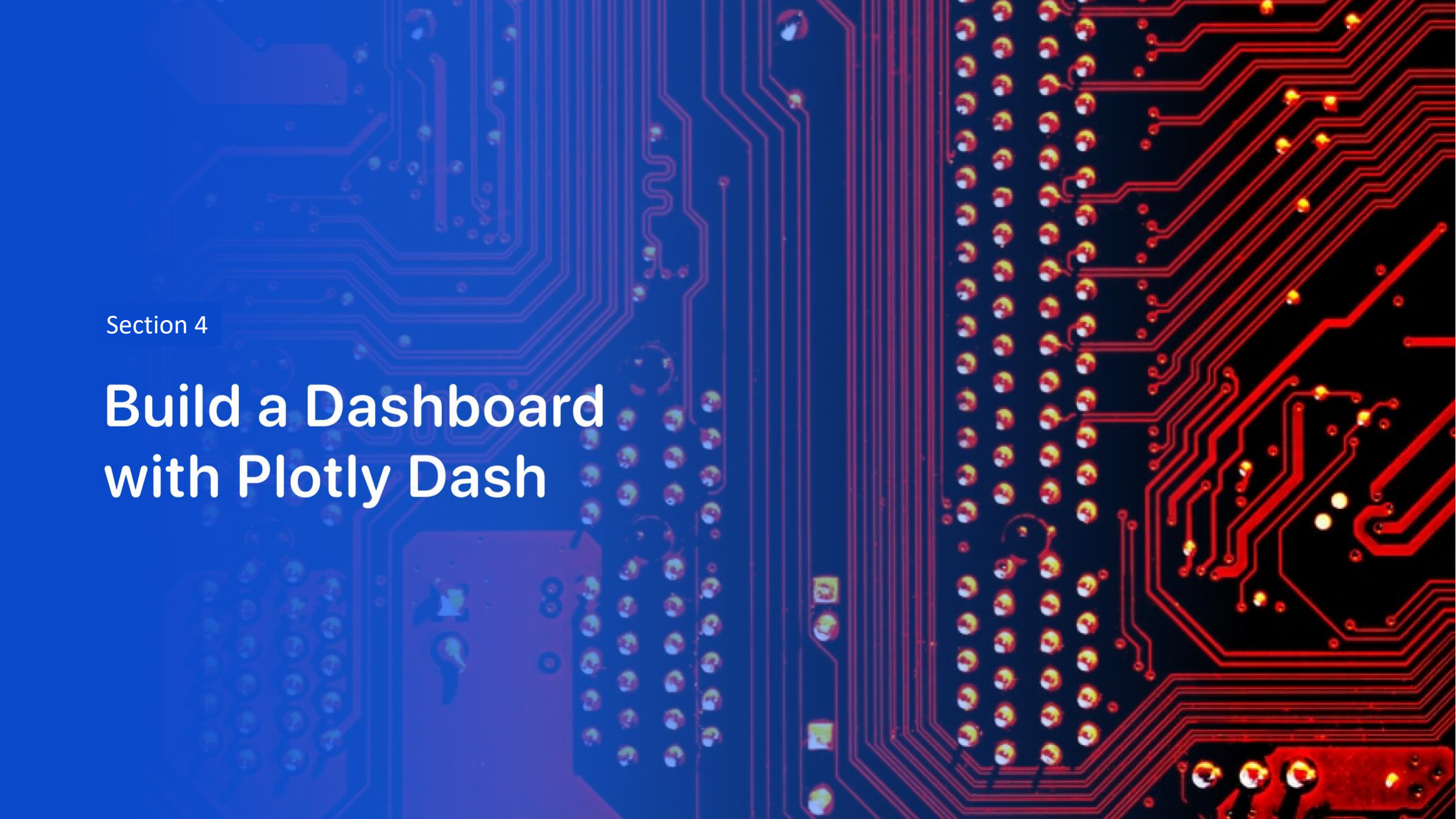
Folium Map – Launch Outcomes



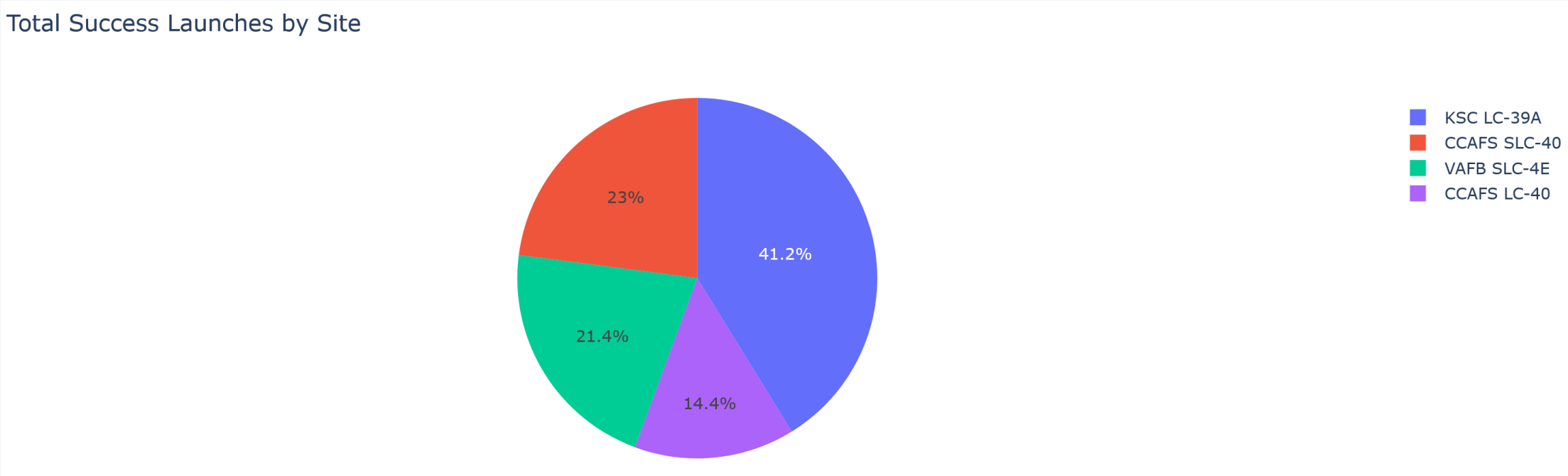
- There were more launches in Florida than in California
- The green markers show successful launches. Red markers show failed launches
- The highest success rate has the launch site: KSC LC-39A

Folium Map - Proximities

- Launch site “CCAFS SLC 40” was chosen for the calculation for distances to proximities.
- Distance to:
  - Coast = 0,88 km
  - Railway = 1,23 km
  - Highway = 6,6 km
  - City = 18.17 km

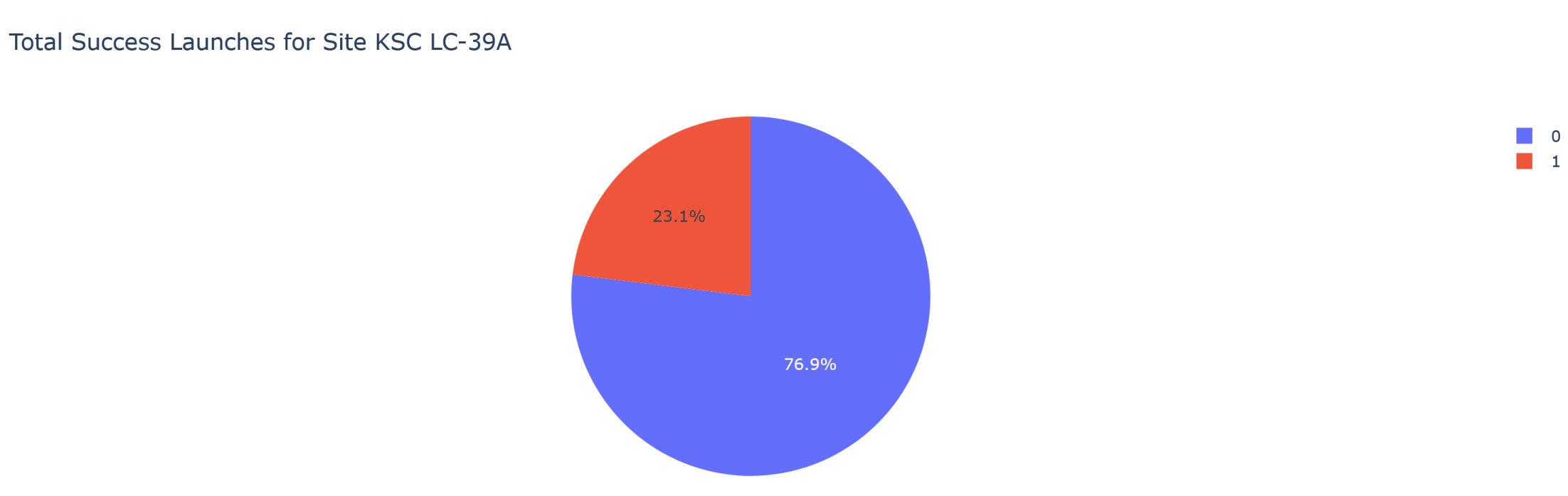

Section 4
Build a Dashboard with Plotly Dash

Dash – Successful Launches per Site



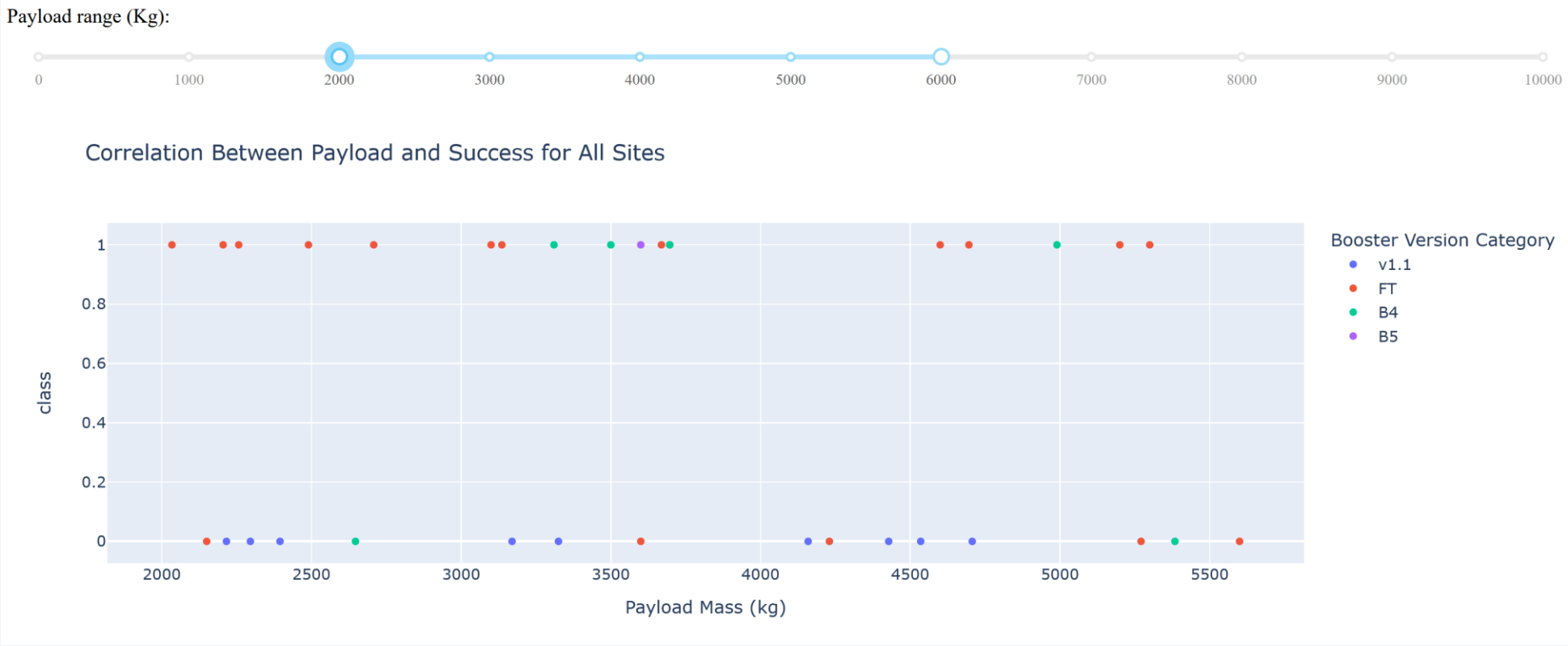
- Launch site KSC LC-39A has not only the highest success rate for launches it has also most of successful launches (41.2 %). It is followed by CCAFS SLC-40 with 23 %.

Dash – Highest Success Rate for Launches



- Launch site KSC LC-39A has the highest success rate for launches (76.9 %).

Dash – Payload and Mission Success



- Payloads between 2000 kg and 5200 kg seem to have the highest rate of success over all sites. Starting with payloads above 5300 kg, the success rate drops significantly.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

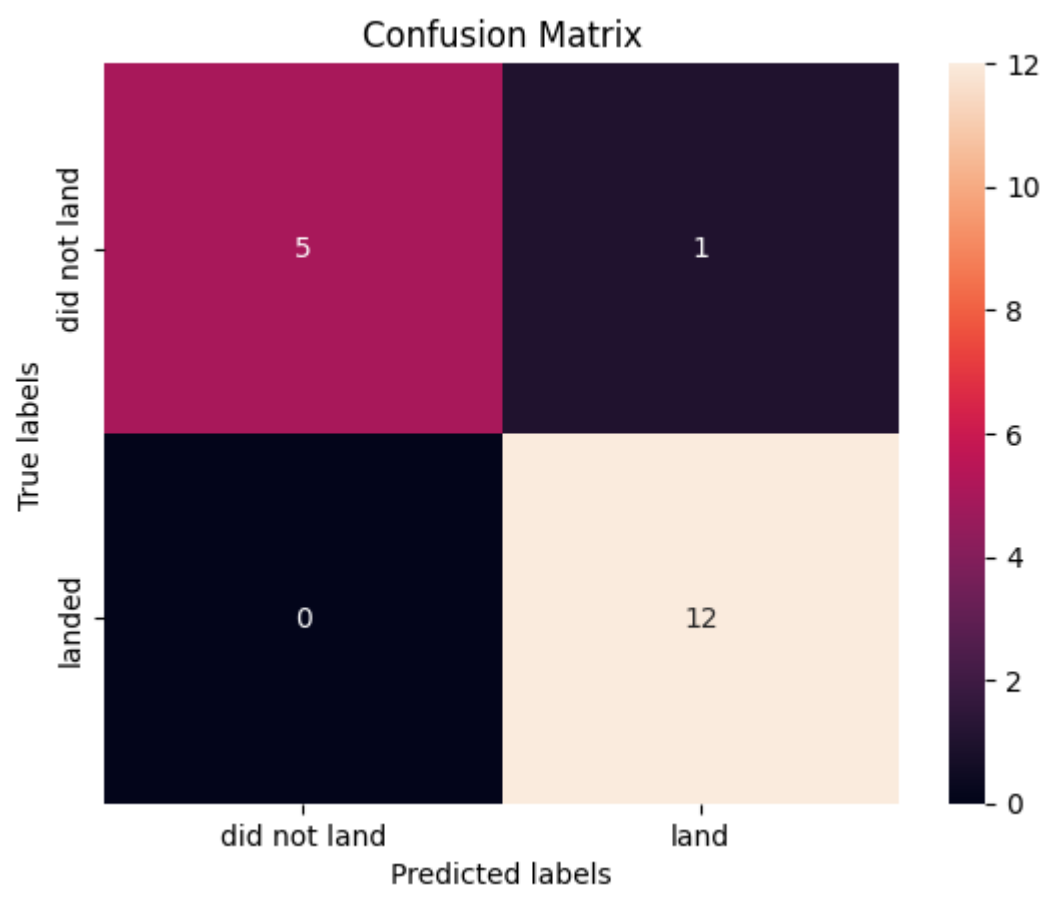
```
methods = {'KNeighbors':KNN_CV.best_score_,
           'DecisionTree':tree_cv.best_score_,
           'LogisticRegression':logreg_cv.best_score_,
           'SupportVector': svm_cv.best_score_}

bestmethod = max(methods, key=methods.get)
print('Method with best performance is', bestmethod,'with a score of', methods[bestmethod])
```

```
Method with best performance is DecisionTree with a score of 0.875
```

- The decision tree model was identified as the model for the predictive analysis with a score of 87.5 %.

Confusion Matrix



- The decision tree model was identified as the model for the predictive analysis
- Data accuracy on test data is 94.4 %
- There is one false positive in the confusion matrix. It means that a successful landing is predicted, but in reality the rocket crashes. This is the most dangerous outcome in the confusion matrix.

Conclusions

- SpaceX has that the commercialized cost-efficient rocket launches can be successful. The rate of successful missions has increased over the years and is very likely to increase even higher.
- Launch sites are located close to the coast and with keeping distance to cities and highways. With a further increasing success rate, it can be assumed that rocket launches from cities might be an option for super fast travel across the globe.
- Higher payloads lead to a decreasing rate of the mission success. Building larger infrastructures in the orbits or on different planets remain a vision or must come along with more launches.
- Launches aiming at Orbits ES-L1, GEO, HEO and SSO had a success rate of 100 %
- The decision tree is the best predictive model, but false positive results in the confusion matrix are a problem and need to be reduced/eliminated.
- It is recommended to extend the analysis on a broader data set.

Thank you!