IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Matej Salamunic
06/05/2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection using Space X API and WebScraping
  - Data wrangling with Falcon 9 rocket data separation and creating landing outcome indicator
  - Exploratory Data Analysis (EDA) with SQL queries
  - Analysis with interactive Folium map and Plotly Dash dashboard
  - Predictive analysis using classification models
- Summary of all results
  - The best resulting models, Decision Tree, have over 88.8% accuracy in predicting whether the Falcon 9 rocket will fail or succeed

Introduction

- Project background and context
  - SpaceX advertises the launch of the Falcon 9 rocket with a price tag of $62 million, while other service providers cost more than $165 million. Much of SpaceX's savings is in reusing the first stage. Based on available information and machine learning models we will predict if SpaceX use the first stage again and then we can have better estimate of the launch cost.
- Problems you want to find answers
  - Which of the variables, such as launch site, payload mass, and to what extent affect whether a rocket will land successfully
  - The best algorithm for predicting whether the Falcon 9 first stage will land successfully

Section 1

Methodology

Methodology

Executive Summary

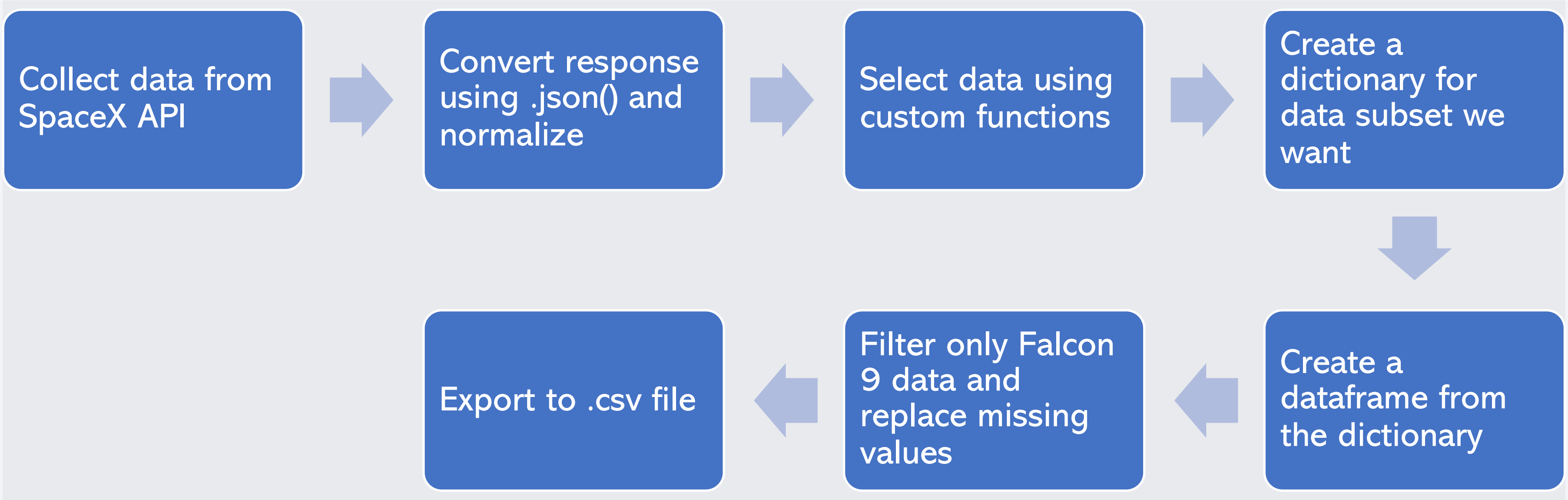
- Data collection methodology:
  - Data was collected from Spacex API and Falcon 9 Wikipedia page web scraping
- Perform data wrangling
  - Filter Falcon 9 rocket data, replacing missing values, creating a landing success outcome label
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The data is normalized and split into train and test datasets
  - Then the models are built and tuned, and the accuracy is calculated to find the best result

Data Collection

- Data sets were collected from two sources using
  - API requests from SpaceX REST API (https://api.spacexdata.com/v4/rockets/)
  - Web Scraping from SpaceX's Wikipedia data (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- Collected data provides information about the take-off date, location, rockets, payload, flight characteristics, landing results, etc (e.g. FlightNumber, Date, Serial, LaunchSite, BoosterVersion, PayloadMass, Orbit, Outcome, Reused, Longitude, Latitude)

Data Collection – SpaceX API

- Flowchart for SpaceX API calls data collection



Collect data from SpaceX API

Convert response using .json() and normalize

Select data using custom functions

Create a dictionary for data subset we want

Create a dataframe from the dictionary

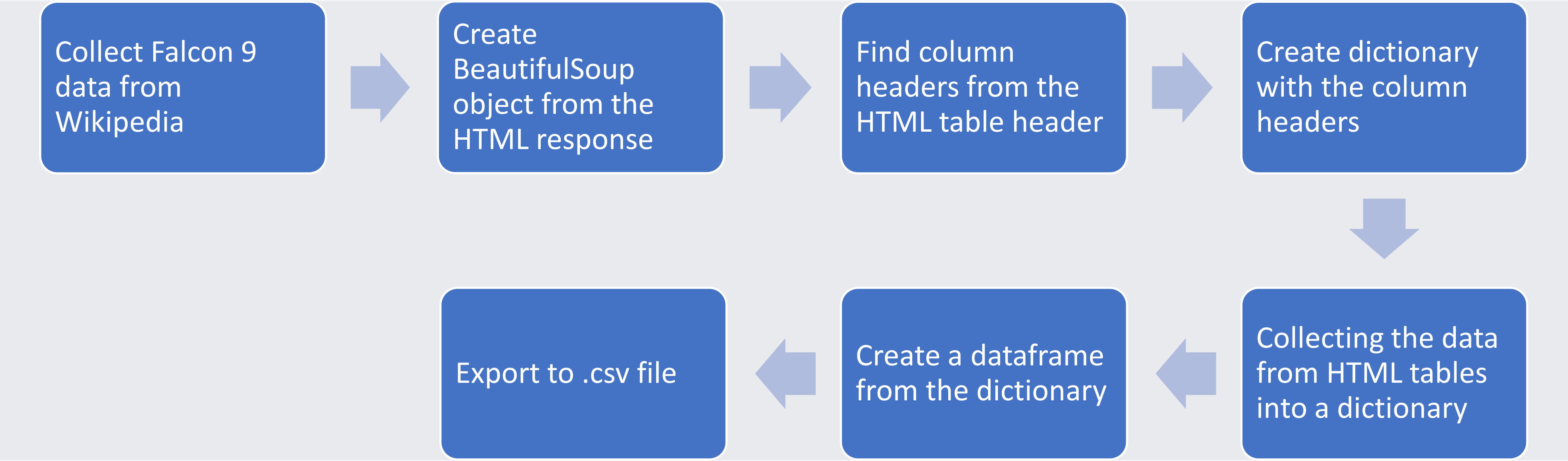
Filter only Falcon 9 data and replace missing values

Export to .csv file

- Source code: https://github.com/matej-s/falcon/blob/master/DC_API.ipynb

Data Collection - Scraping

- Flowchart for data obtained from Wikipedia



Collect Falcon 9 data from Wikipedia

Create BeautifulSoup object from the HTML response

Find column headers from the HTML table header

Create dictionary with the column headers

Collecting the data from HTML tables into a dictionary

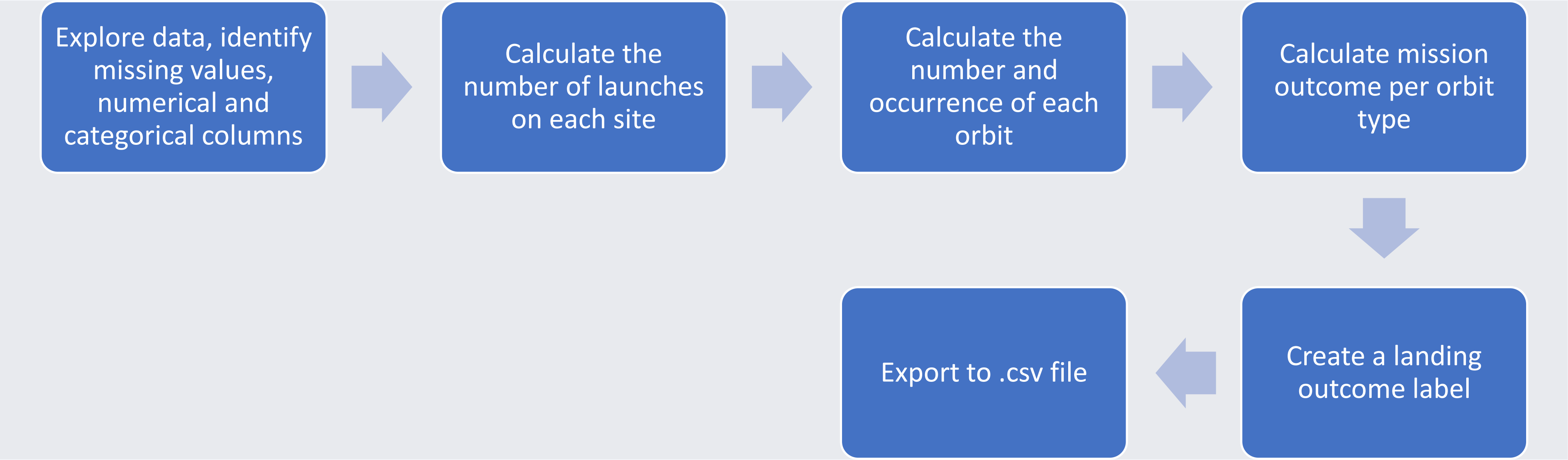
Create a dataframe from the dictionary

Export to .csv file

- Source code: https://github.com/matej-s/falcon/blob/master/DC_Web_Scraping.ipynb

Data Wrangling

- First, an overview of the data set was done to see the percentage of missing values and to identify data types.
- Then the launches per site, occurrences of each orbit and mission outcome per orbit type occurrences were calculated.
- Finally, the landing outcome label was created using Outcome column data.

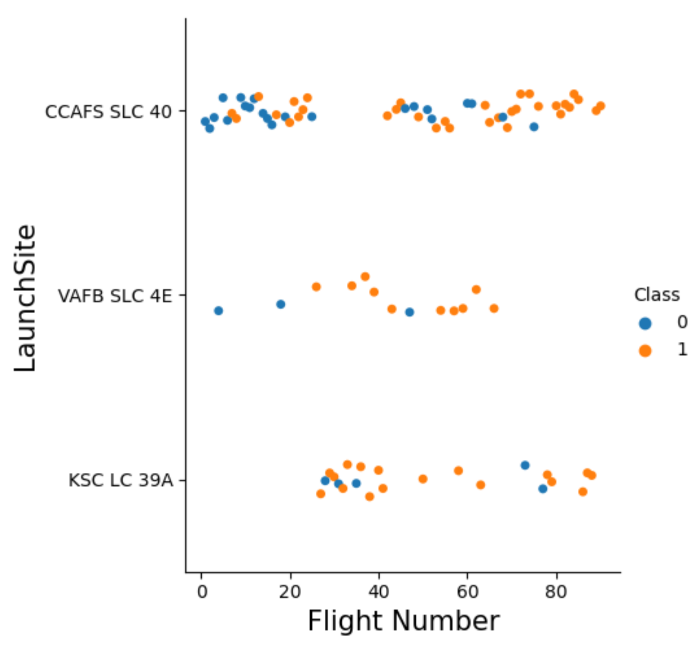

Explore data, identify missing values, numerical and categorical columns
Calculate the number of launches on each site
Calculate the number and occurrence of each orbit
Calculate mission outcome per orbit type
Create a landing outcome label
Export to .csv file

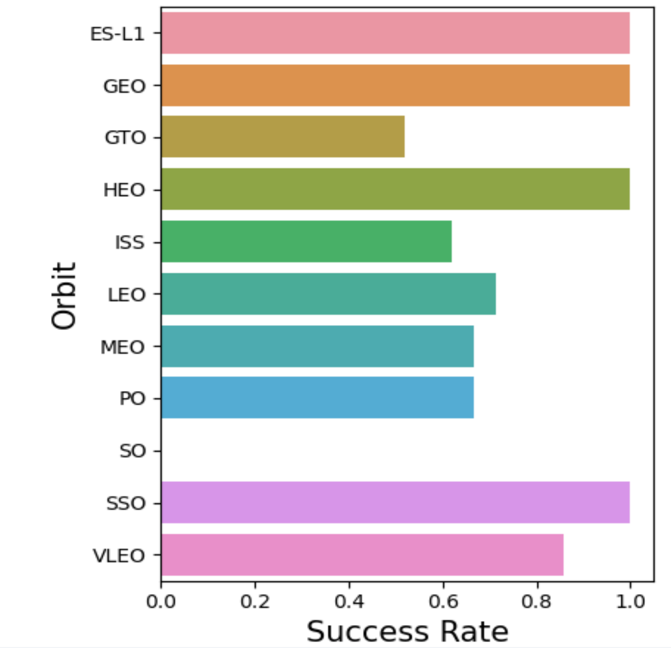
- Source code: https://github.com/matej-s/falcon/blob/master/Data_Wrangling.ipynb

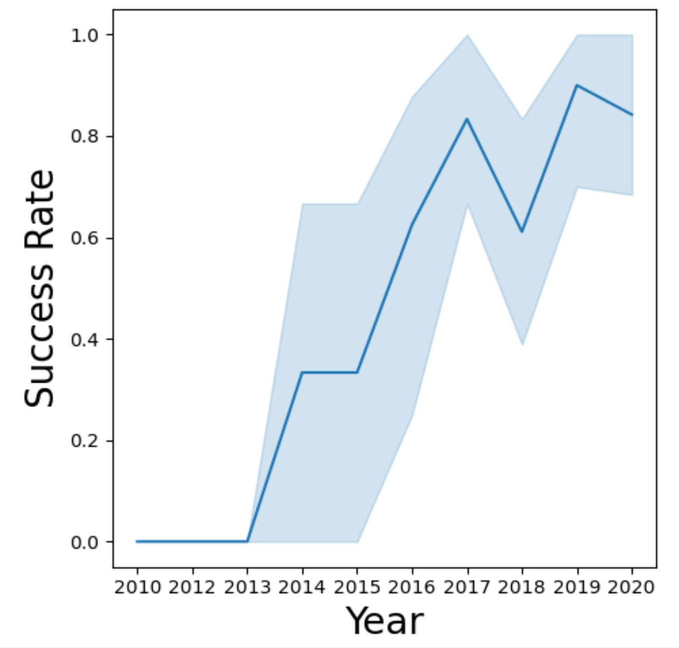
EDA with Data Visualization

- Exploratory Data Analysis (EDA) carried out to investigate relationships between variables FlightNumber, PayloadMass, LaunchSite, Orbit, Class and Year.
- Scatterplots, barplots and line charts were used to visualize the relationship
    - Scatter plots show the relationship between variables so they could be used in machine learning model
    - Bar charts show relationship between categories being compared
    - Line charts show trends in data over time

EDA with Data Visualization







- Based on the charts we obtain insights about how each variable would affect the success rate.

- Source code: https://github.com/matej-s/falcon/blob/master/jupyter-labs-eda-data_vizualization.ipynb

EDA with SQL

- SQL queries performed
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'KSC'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date where the first successful landing outcome in drone ship was achieved
  - List the names of the boosters which have success in ground pad and have payload mass greater than 4.000 but less than 6.000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster versions which have carried the maximum payload mass
  - List the records which will display the month names, successful landing outcomes in ground pad, booster versions, launch site for the months in year 2017
  - Rank the count of successful landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order

- Source code: https://github.com/matej-s/falcon/blob/master/jupyter-labs-EDA-SQL-edx.ipynb

Build an Interactive Map with Folium

- Maps were created using markers, circles, marker clusters and lines
    - Markers with circle using coordinates are used to geographical show launch locations
    - Marker cluster of success and failed launches are used to identify launch sites with high success rates
    - Lines are used to show the distance between launch sites and the closest city, highway, coast, railway
- Using an interactive map of launch locations, we look for geographic patterns and factors for an optimal launch sites

- Source code: https://github.com/matej-s/falcon/blob/master/jupyter_lab_Visual_Analytics_with_Folium.ipynb

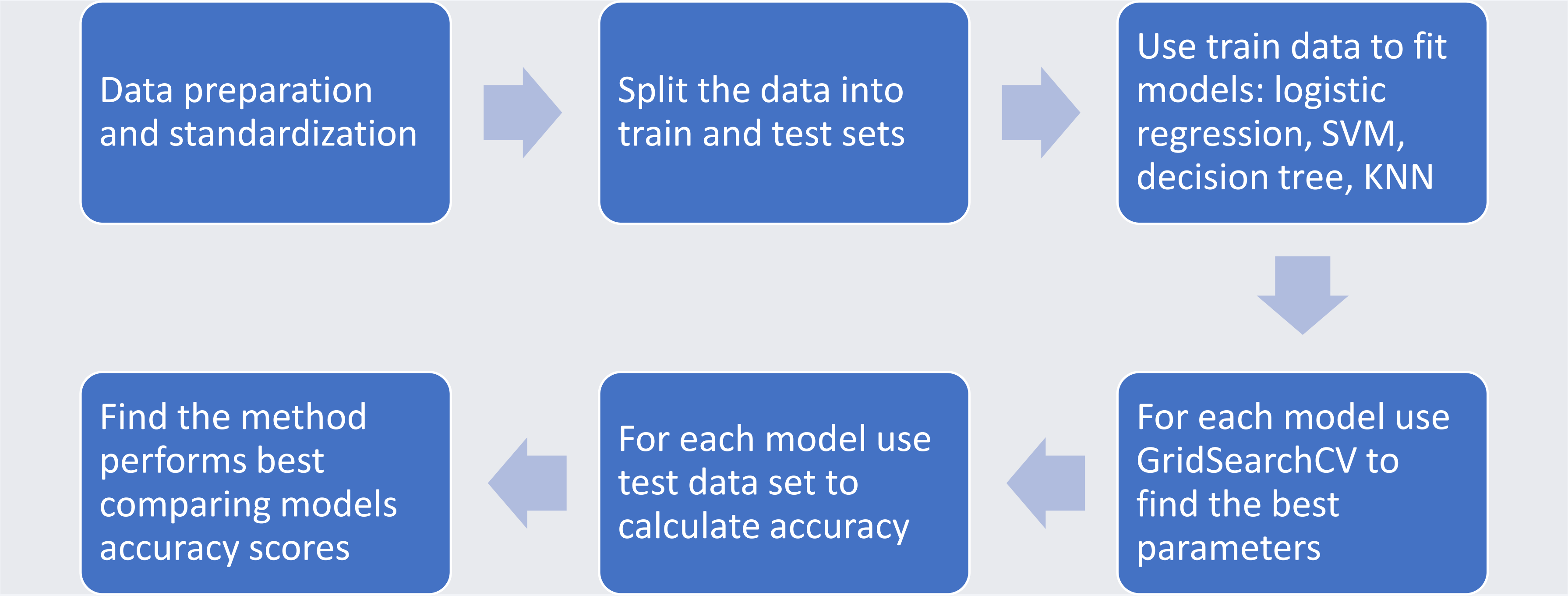
Build a Dashboard with Plotly Dash

- Interactive Dashboard with a pie chart and a scatter plot was used to analyze the relation between launch sites and payloads
  - Dropdown list enable Launch Site selection
  - Pie chart show successful launch count by sites and success failed ratio for selected site
  - Slider allows selection of the Payload Mass range
  - Scatter chart of Payload and Launch Success for the different Booster Versions show the correlation between Payload and Launch Succes

- Source code: https://github.com/matej-s/falcon/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

- Four classification models were compared
  - Logistic regression, support vector machine (SVM), decision tree and k nearest neighbors (KNN)



Data preparation and standardization

Split the data into train and test sets

Use train data to fit models: logistic regression, SVM, decision tree, KNN

For each model use GridSearchCV to find the best parameters

For each model use test data set to calculate accuracy

Find the method performs best comparing models accuracy scores

Source code: https://github.com/matej-s/falcon/blob/master/Machine_Learning_Prediction_Part_5_2.jupyterlite.ipynb

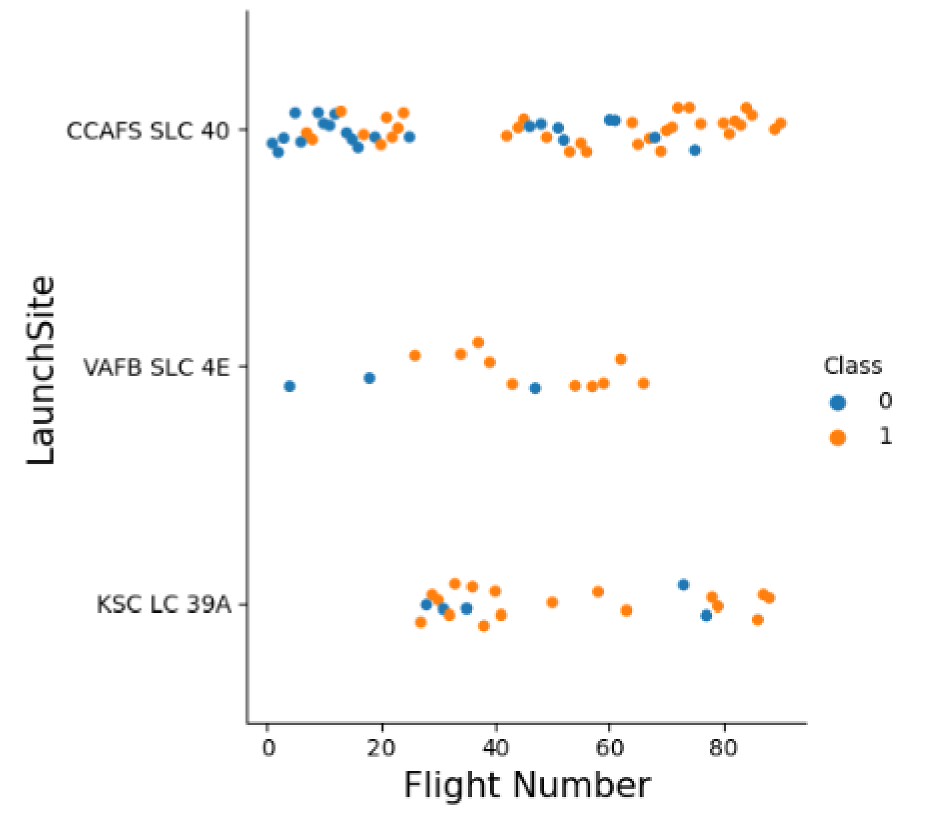
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

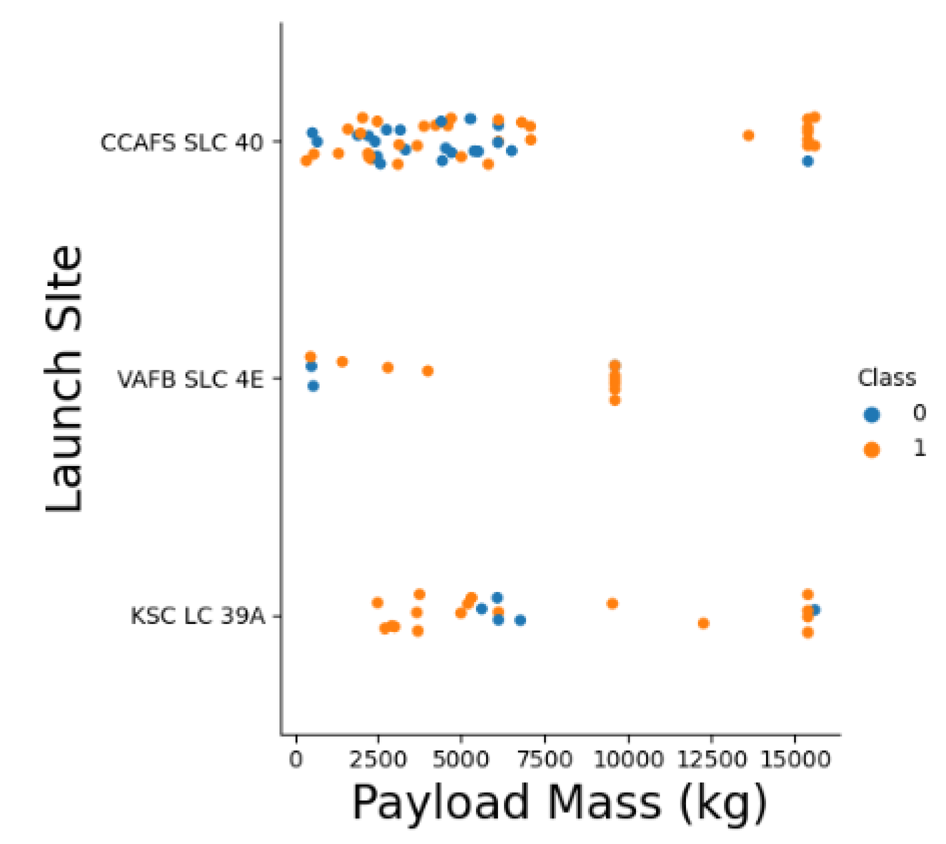
Flight Number vs. Launch Site



- First flights have more unsuccessful launches while recent flights have more success

- The launch site CCAFS SLC 40 has a significantly higher number of launches than the launch sites VAFB SLC 4E and KSC LC 39A

Payload vs. Launch Site

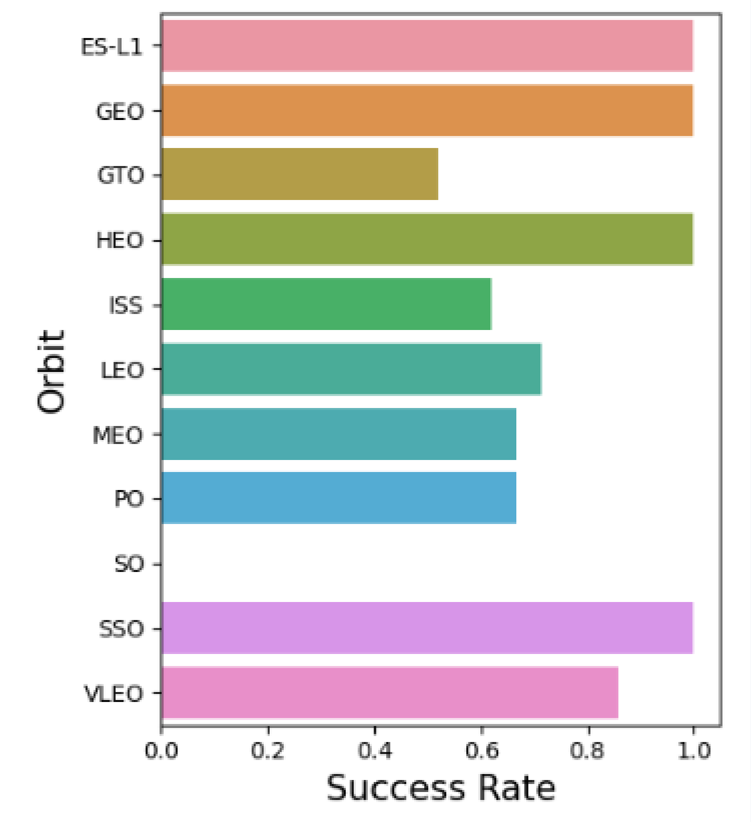


- The KSC LC 39A has a very high success rate for Payload Mass below 5.000 and above 10.000

- The VAFB SLC 4E site has a very high success rate for Payload Mass above 2.500

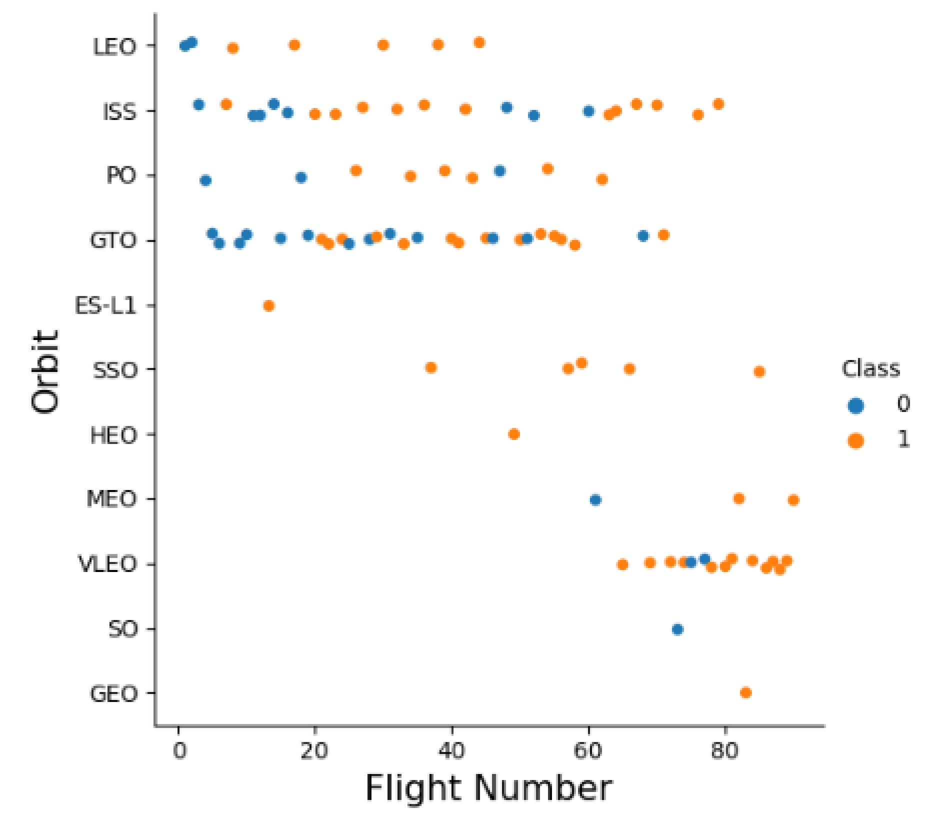
- In general most launches with Payload Mass above 7.500 were successful

Success Rate vs. Orbit Type



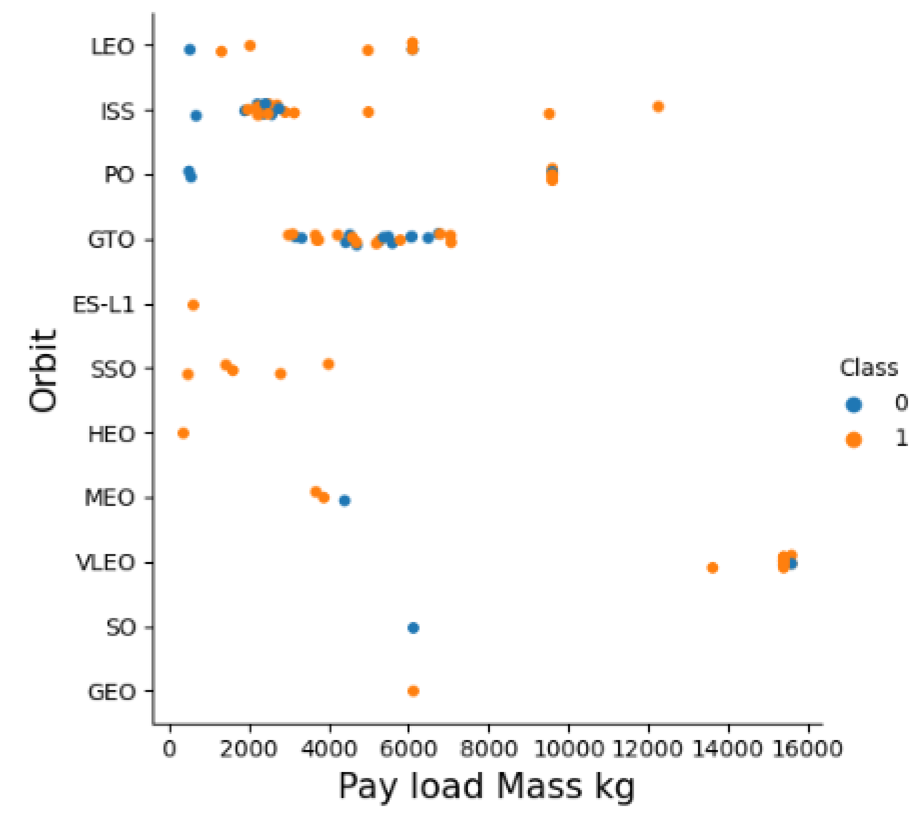
- Four orbits have 100% success launch rate
- Five orbits have success launch rate between 50% and 85%
- One orbit have 0% successful launch rate

Flight Number vs. Orbit Type



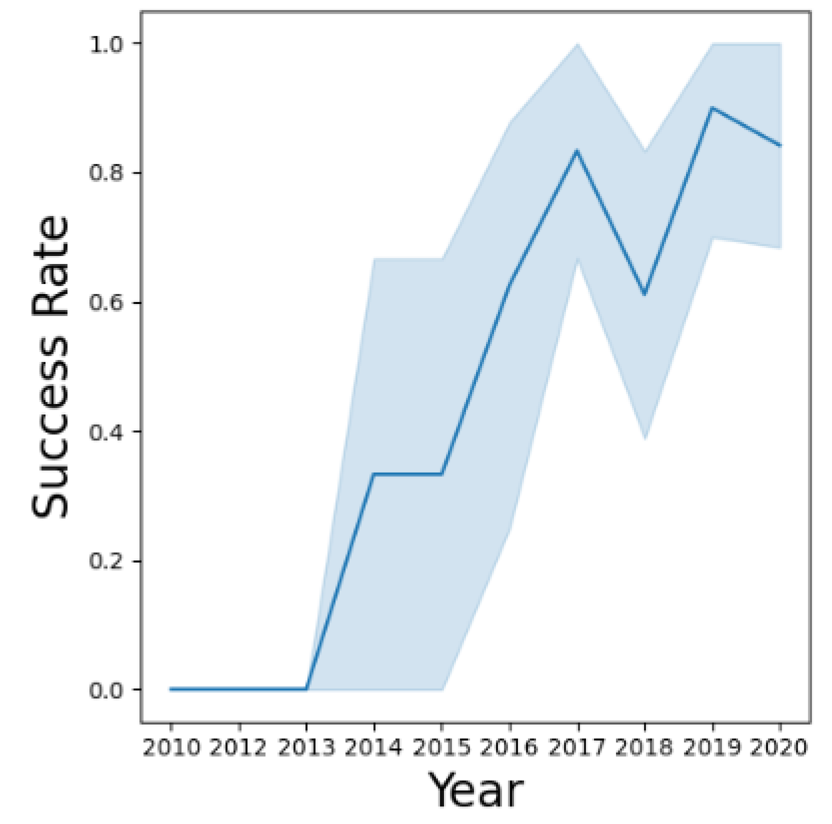
- LEO orbit success seems to be related to the number of flights, first failed and then successful launches
- SSO orbit has a high success rate in general
- There is no correlation between GTO orbit and flight number

Payload vs. Orbit Type



- PO, ISS and LEO orbit have more successful landing with heavy payloads
- GTO orbit and Payload Mass have no correlation
- Most launches are with a Payload Mass below 10.000

Launch Success Yearly Trend

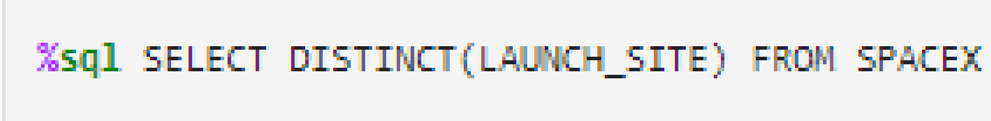


- The success rate is increasing from 2013 to 2020
  - except for 2018 when the success rate decreased

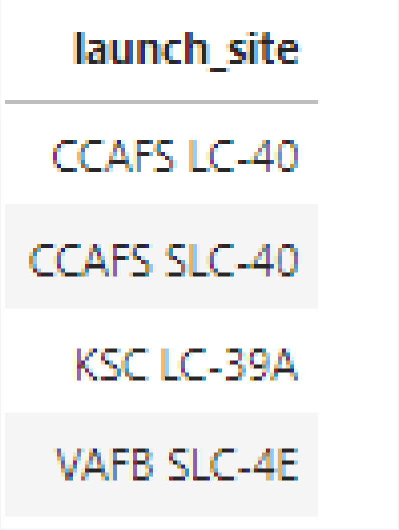
- The highest success rate was recorded in 2019

All Launch Site Names

- Query for unique launch site name



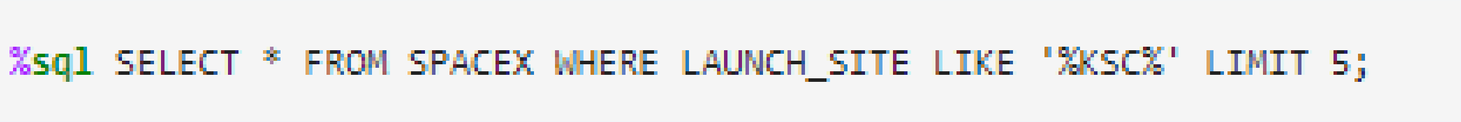
```
%sql SELECT DISTINCT(LAUNCH_SITE) FROM SPACEX
```



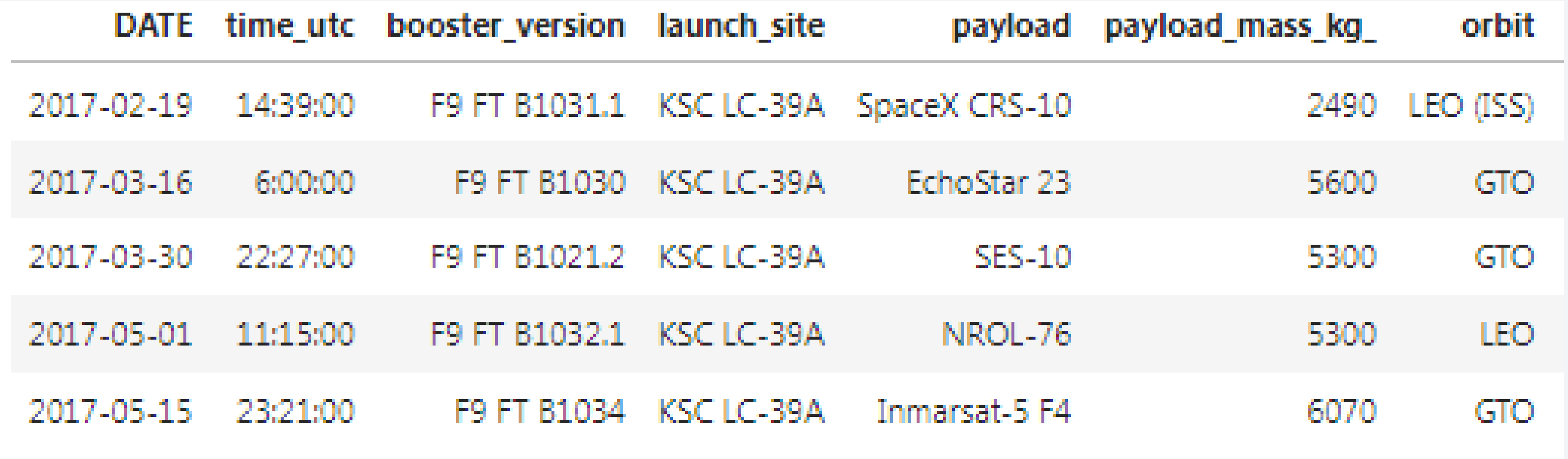
launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Launch Site Names Begin with 'KSC'

- Query for the first five records in the database where launch site name begins with ‘KSC’



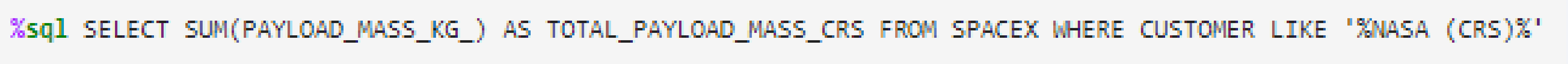
```
%sql SELECT * FROM SPACEX WHERE LAUNCH_SITE LIKE '%KSC%' LIMIT 5;
```



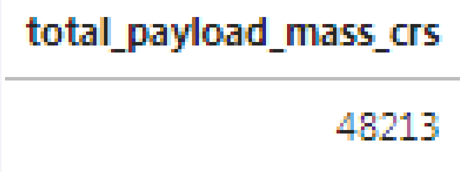
DATE	time_utc	booster_version	launch_site	payload	payload_mass_kg_	orbit
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)
2017-03-16	6:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO
2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO

Total Payload Mass

- SQL for the total payload mass carried by boosters launched by NASA (CRS)



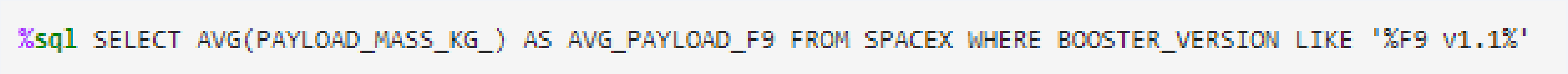
```
%sql SELECT SUM(PAYLOAD_MASS_KG_) AS TOTAL_PAYLOAD_MASS_CRS FROM SPACEX WHERE CUSTOMER LIKE '%NASA (CRS)%'
```



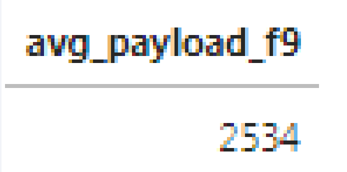
total_payload_mass_crs
48213

Average Payload Mass by F9 v1.1

- SQL for the average payload mass carried by booster version F9 v1.1



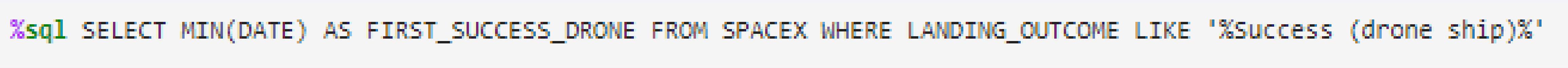
```
%sql SELECT AVG(PAYLOAD_MASS_KG_) AS AVG_PAYLOAD_F9 FROM SPACEX WHERE BOOSTER_VERSION LIKE '%F9 v1.1%'
```



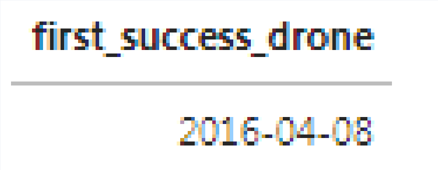
avg_payload_f9
2534

First Successful Ground Landing Date

- SQL for the first successful landing outcome



```
%sql SELECT MIN(DATE) AS FIRST_SUCCESS_DRONE FROM SPACEX WHERE LANDING_OUTCOME LIKE '%Success (drone ship)%'
```

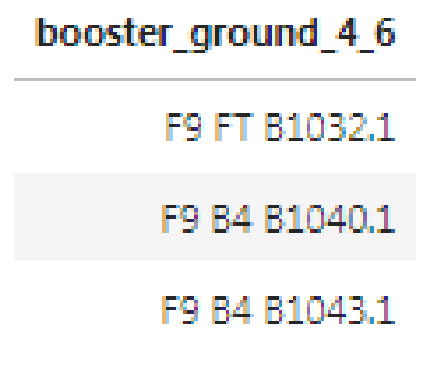


first_success_drone
2016-04-08

Successful Drone Ship Landing with Payload between 4000 and 6000

- SQL for the first successful landing outcome date, payload between 4000 and 6000

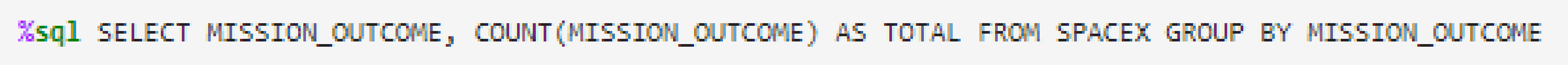
  - %sql SELECT BOOSTER_VERSION AS BOOSTER_GROUND_4_6 FROM SPACEX WHERE (PAYLOAD_MASS_KG_ BETWEEN 4001 AND 5999) AND LANDING_OUTCOME LIKE '%Success (ground pad)%'



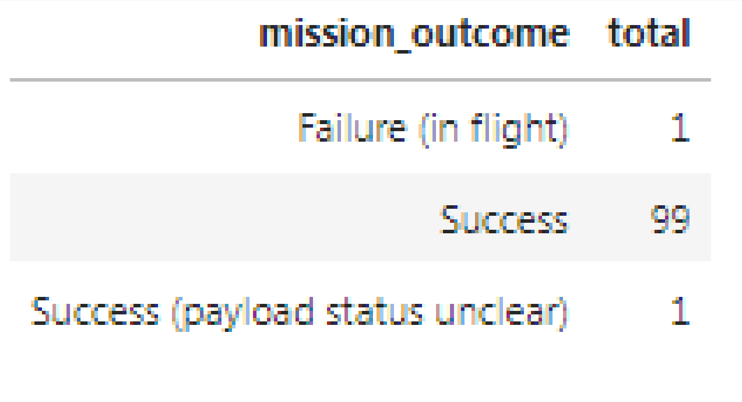
booster_ground_4_6
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

Total Number of Successful and Failure Mission Outcomes

- SQL for the total number of successful and failure mission outcomes



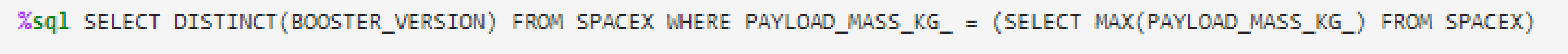
```
%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS TOTAL FROM SPACEX GROUP BY MISSION_OUTCOME
```



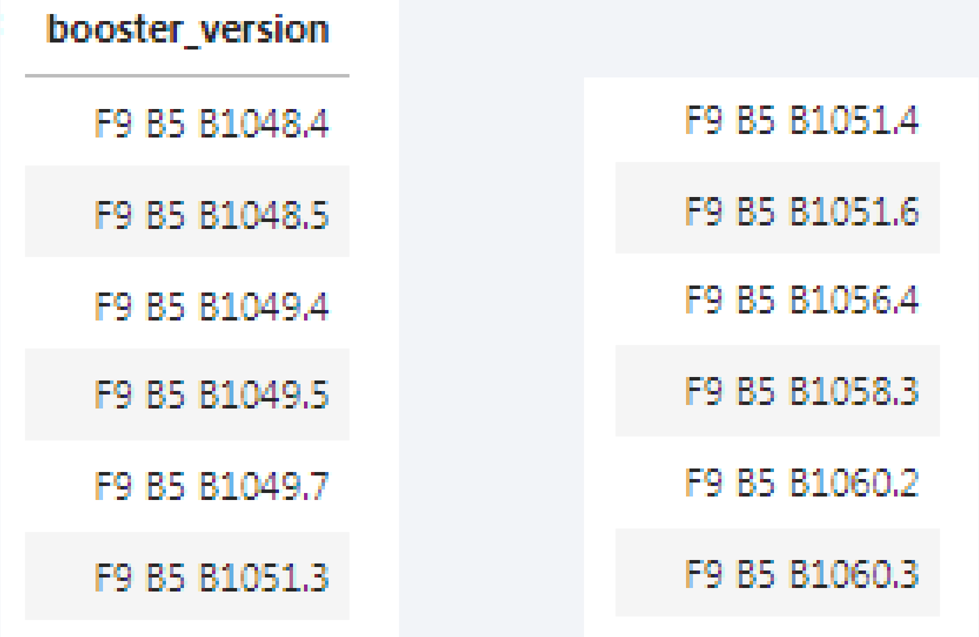
mission_outcome	total
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- SQL for the booster which have carried the maximum payload mass



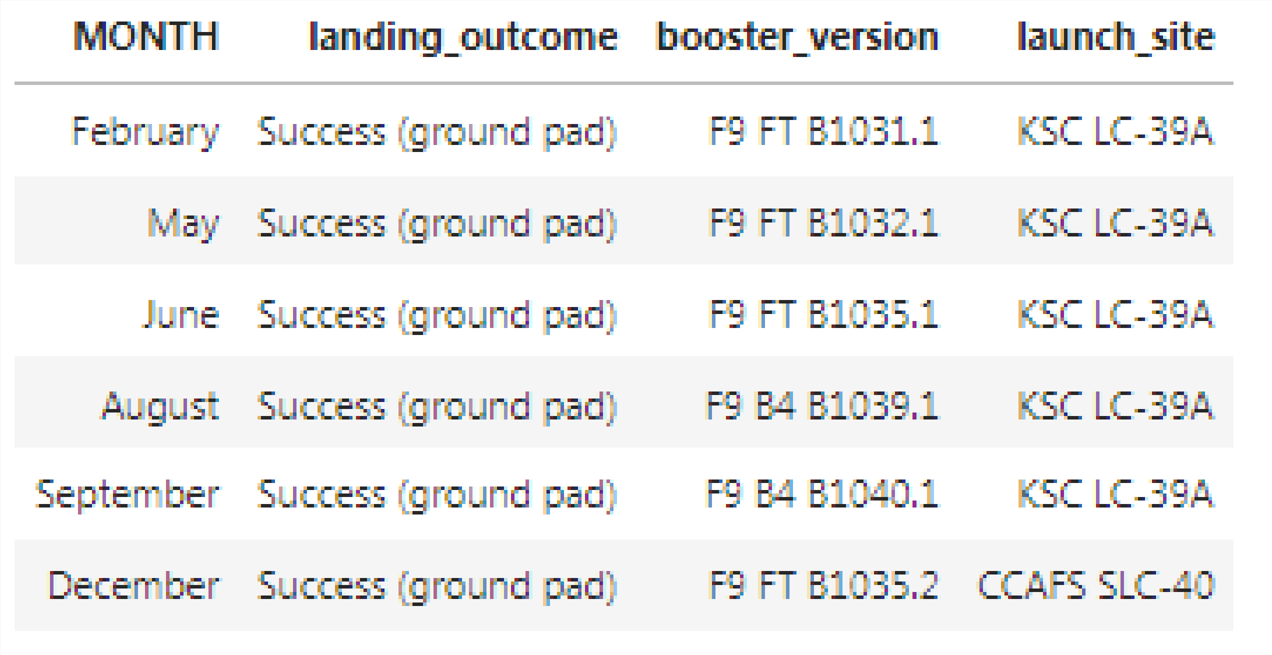
```
%sql SELECT DISTINCT(BOOSTER_VERSION) FROM SPACEX WHERE PAYLOAD_MASS_KG_ = (SELECT MAX(PAYLOAD_MASS_KG_) FROM SPACEX)
```



booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2017 Launch Records

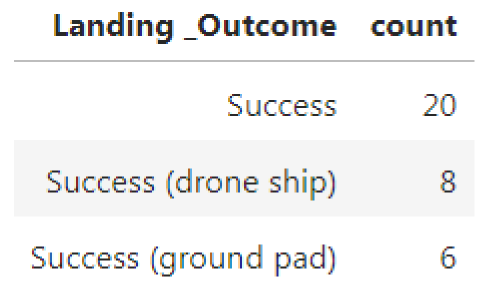
- SQL for months in 2017 with successful landing outcomes
  - %sql SELECT MONTHNAME(DATE) AS MONTH, LANDING_OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEX WHERE (YEAR(DATE) = 2017 AND LANDING_OUTCOME LIKE '%Success (ground pad)%')



MONTH	landing_outcome	booster_version	launch_site
February	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
May	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
June	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
August	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
September	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
December	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

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

- Sql for the count of successful landing outcomes between 2010-06-04 and 2017-03-20
  - %sql SELECT [Landing _Outcome], count(*) as count from SPACEXTBL
    where ([Landing _Outcome] LIKE ‘%success%' and (DATE BETWEEN '04-06-2010' AND '20-03-2017'))
    Group by [Landing _Outcome] ORDER By count Desc



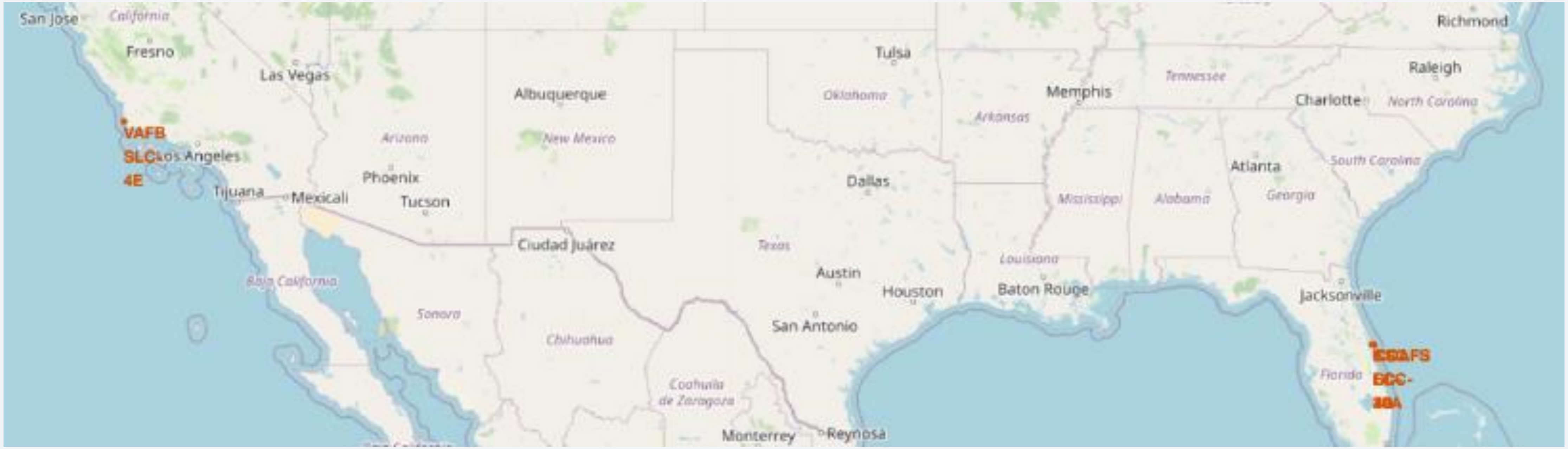
Landing _Outcome	count
Success	20
Success (drone ship)	8
Success (ground pad)	6

Section 3

Launch Sites Proximities Analysis

Launch site locations

- Generated folium map with location markers of all launch sites on a global map

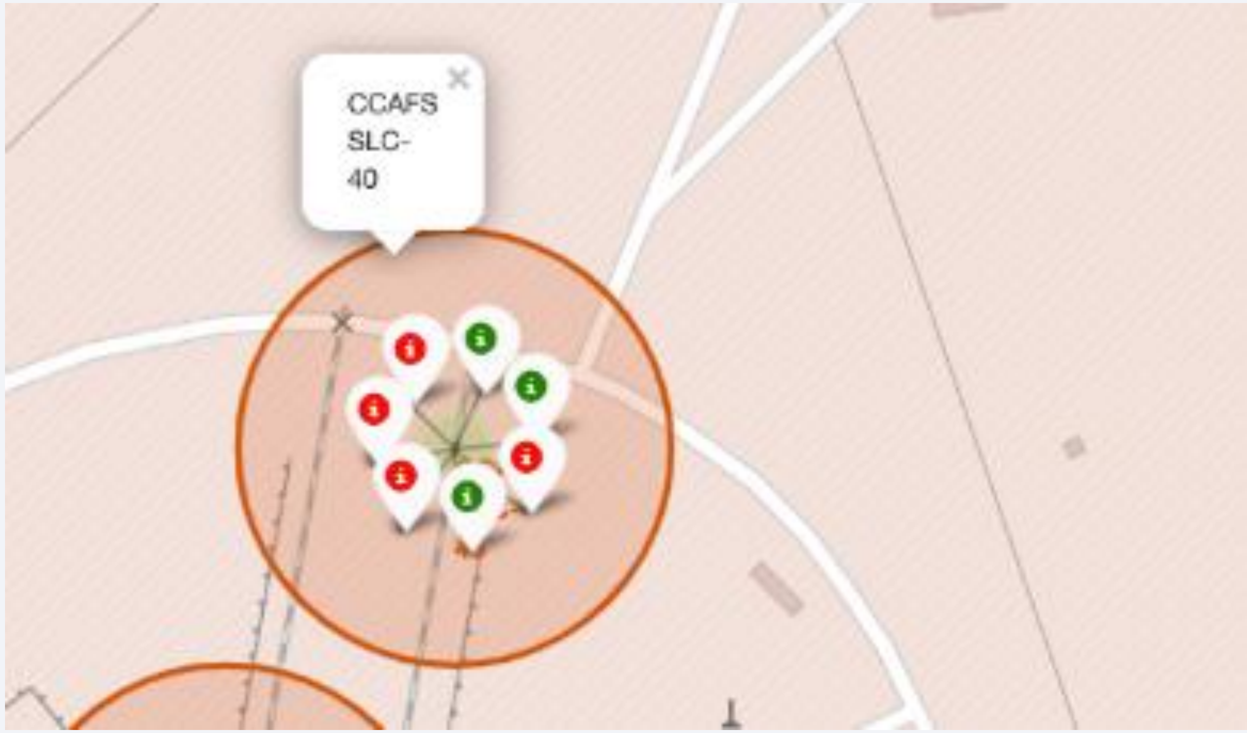


Launch sites are

- located in North America
- located near to coastlines
- in proximity to the Equator line

Launch outcomes marked by color

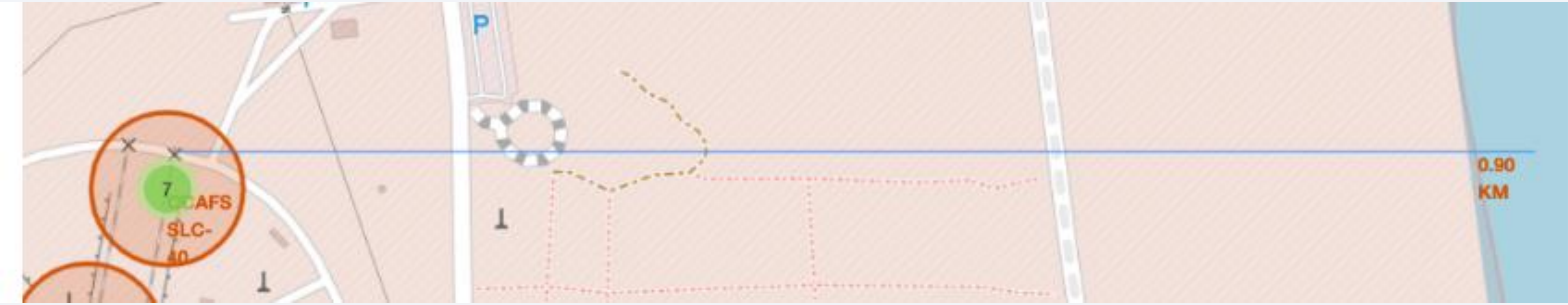
- Success/failed launches for each site are marked on the map
  - green marker is successful launch
  - red marker is unsuccessful launch



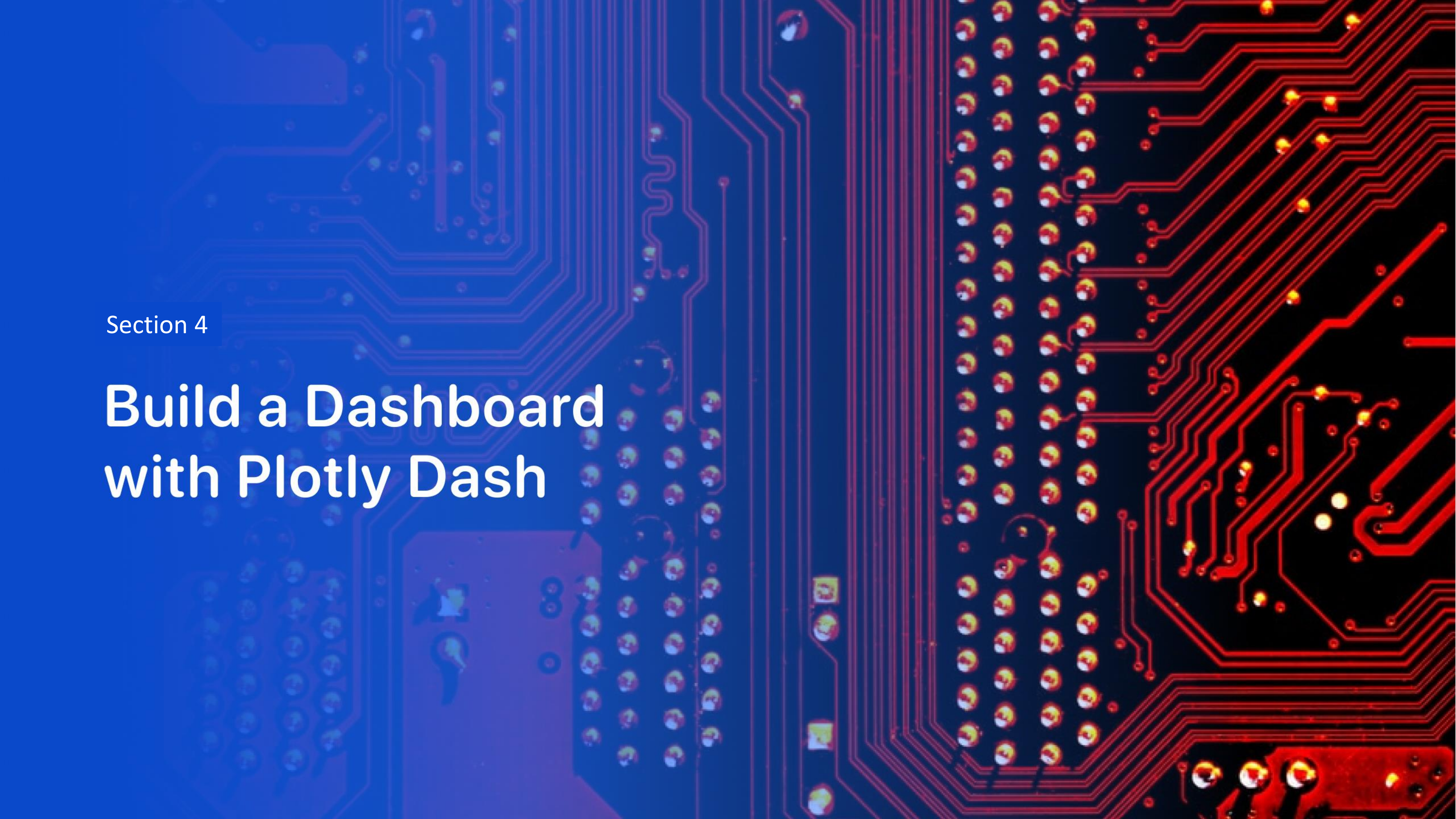
- Color-labeled markers in marker clusters help identify launch sites with a high success rate

Distances between the launch site and nearby locations

- Ddistances between a launch site to its proximities is calculated



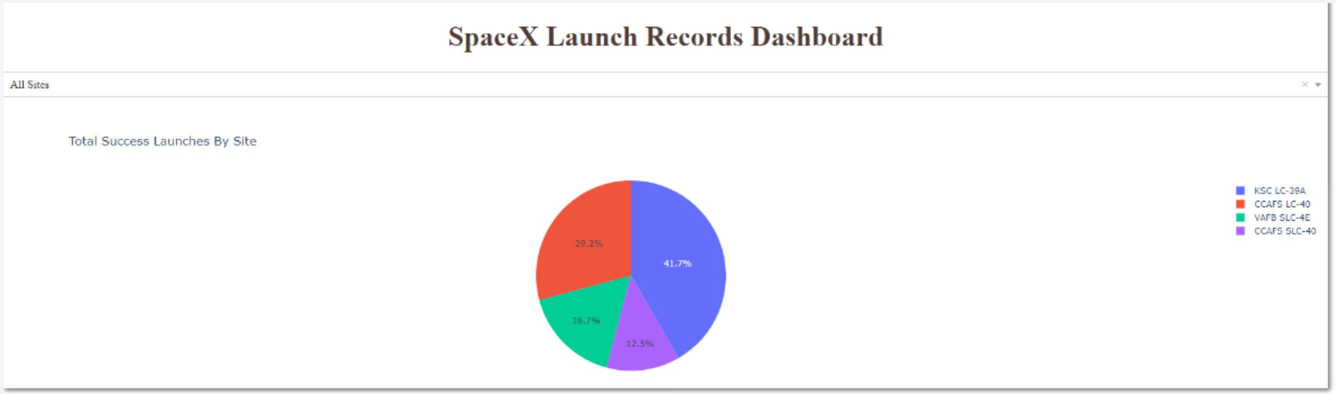
- From the visual analysis of the launch site (CAFS SLC-40) we can see that it is close to coast line (0.9 km)



Section 4

Build a Dashboard with Plotly Dash

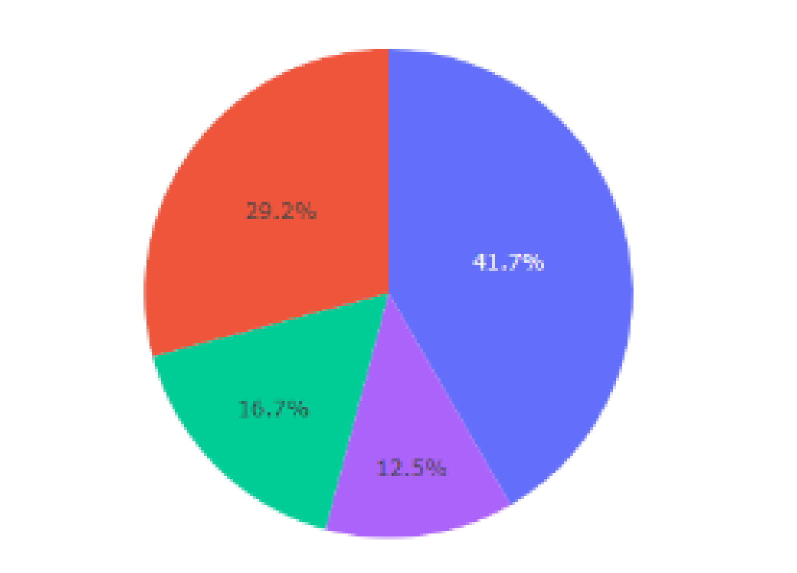
Total Successful Landings by Launch Site

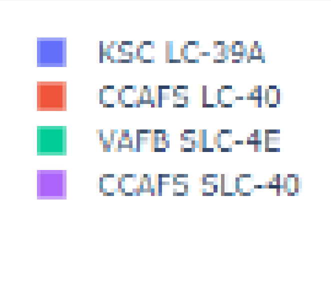


SpaceX Launch Records Dashboard

All Sites

Total Success Launches By Site

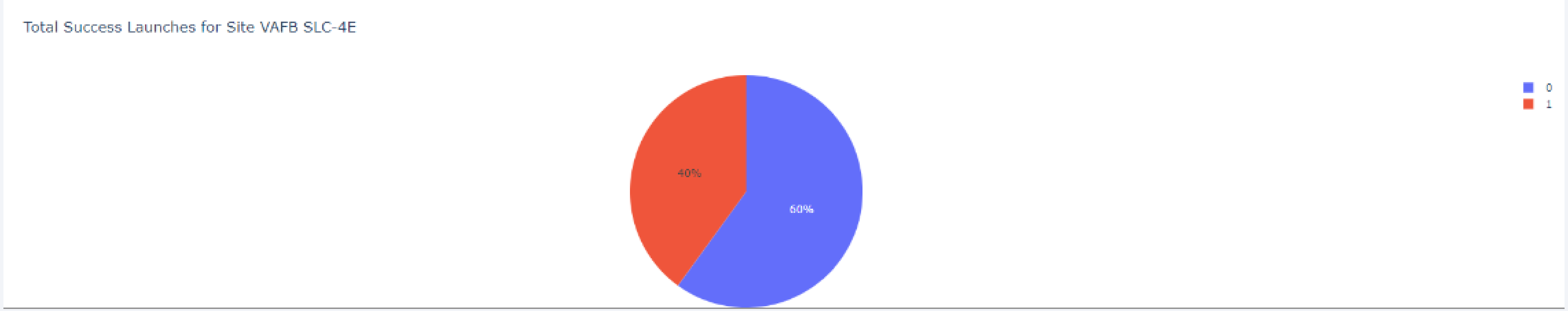




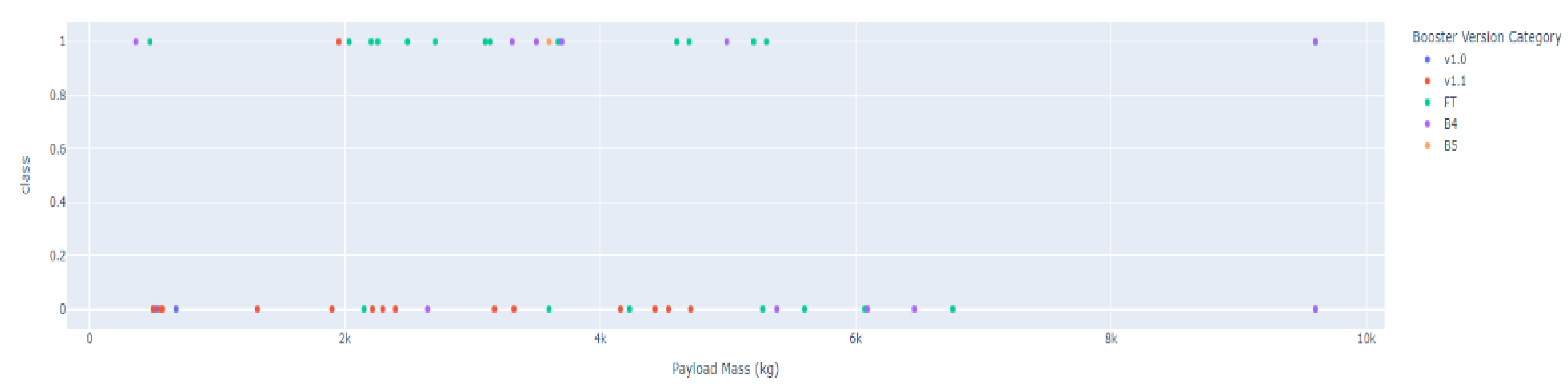
- The most successful landings were launches from KSC LC-39A site
- The least successful landings were launches from CCAFS SLC-40 site

Launch site success ratio

- Launch site locations have different success rates (success vs. failure landings)
- The site with the highest number of successful landings, KSC LC-39A, has 77% of successful landings
  - eg the VAFB SLC-4E location, is the third location in the total number of successful landings. Of the total launch number at this location, 40% are successful and 60% are unsuccessful



Correlation between Payload and Success for all Sites



- the most successful launches have occurred with payloads below 6.000
- the highest launch success rate is for the payload range with 2.000 – 4.000
- FT boosters and payload mass below 6.000 are the most successful combination

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Classification models were built and tested:
  - logistic regression (LogReg)
  - support vector machine (SVM)
  - decision tree classifier (DT)
  - k nearest neighbors (KNN)

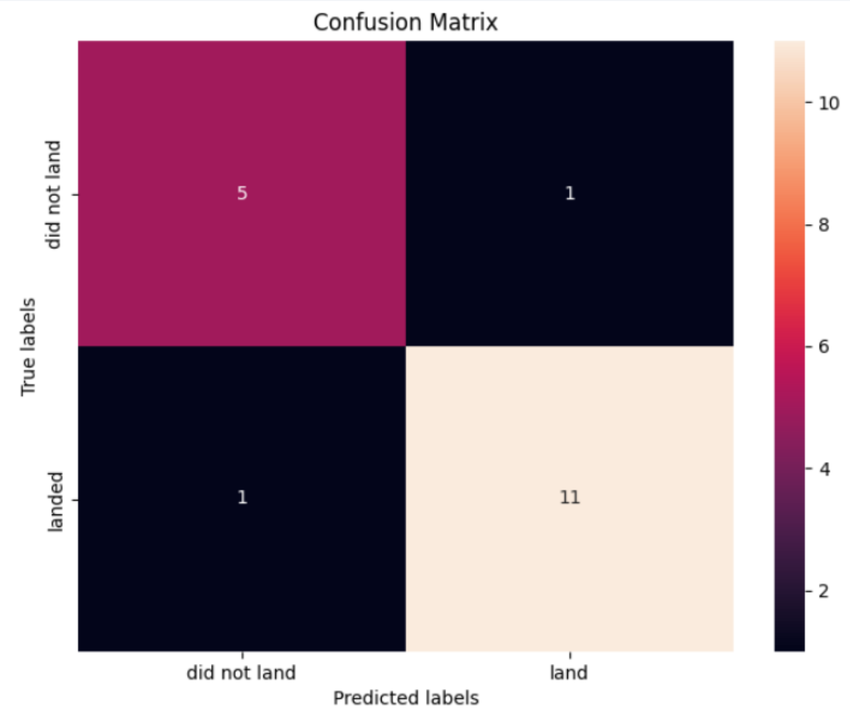
- For each model train data set is used to find the best parameters and then on the test data accuracy is calculated to find the method that performs best

Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
DT	0.875	0.88889
KNN	0.84821	0.83333

- The best model is the Decision Tree Model (DT) with highest accuracy over 88.8%

Confusion Matrix

- Confusion matrix of the best performing model Decision Tree Model (DT)



- Decision Tree confusion matrix show calculated accuracy
    - true positive and true negative compared to false positives and false negatives make high accuracy
    - high score accuracy is made of true positive (11 landed / land) and true negative (5 did not land / did not land) results

Conclusions

- The goal of the project was to develop a machine learning model to predict whether SpaceX will reuse the first stage, and then we can have a better estimate of launch costs

- Conclusions after data analysis
  - successful landing outcomes to reuse the first stage improve over the years
  - KSC LC-39A site has the highest number, 77%, of successful landings
  - ES-L1, GEO, HEO and SSO orbits have 100% success rate
  - low payload mass show better launch results than launches with a larger payload mass

- The best algorithm for predicting first stage successful landings is Decision Tree Classifier with highest accuracy over 88.8%

Appendix

- Project GitHub URL

Thank you!