IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jaan Roos
26 May 2023

Outline

Executive Summary

Introduction

Methodology

Results

Conclusion

Appendix

Executive Summary

Summary of methodologies

- Data collection thtough API
- Data wrangling
- Exploratory Data Analysis with data visualization
- Exploratory Data Analysis with SQL
- Map visualization with Folium
- Interactive dashboard
- Machine Learning/ Predictive analysis

Summary of all results

- Exploratory Data Analysis results
- Interactive analytics results
- Predictive analysis results

Introduction

Project background and context

SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars. Other providers estimate cost of upward of 165 million dollars. SpaceX savings are because it can reuse the first stage launcher. SpaceX information can be used to analyze if competition can be provided for SpaceX launches. The goal of the project is to create machine learning pipelines to predict first stage successful landings, reusability and thus cost savings.

Problems you want to find answers

What factors affect successful landings?

Section 1

Methodology

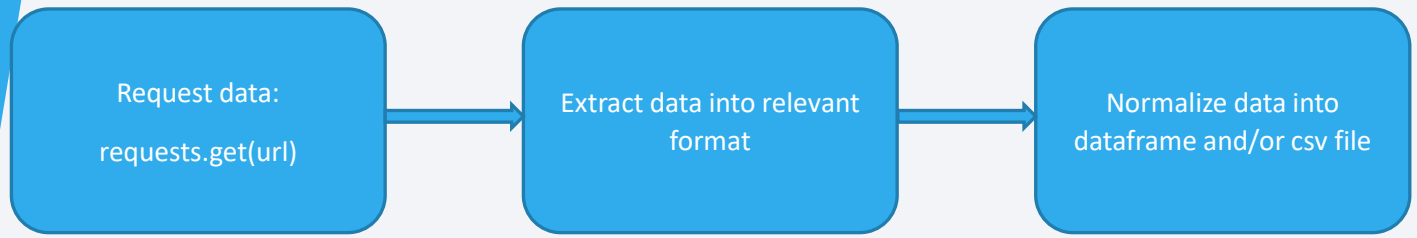
Methodology

Executive Summary

- Data collection methodology:
  - SpaceXAPI
  - WebScraping (Wikipedia)
- Perform data wrangling
  - Cleaning unnecessary data (Falcon 1 rows, irrelevant columns)
  - Creating outcome labels
- 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 normalized, training and test sets created, 4 classification models used and evaluated

Data Collection

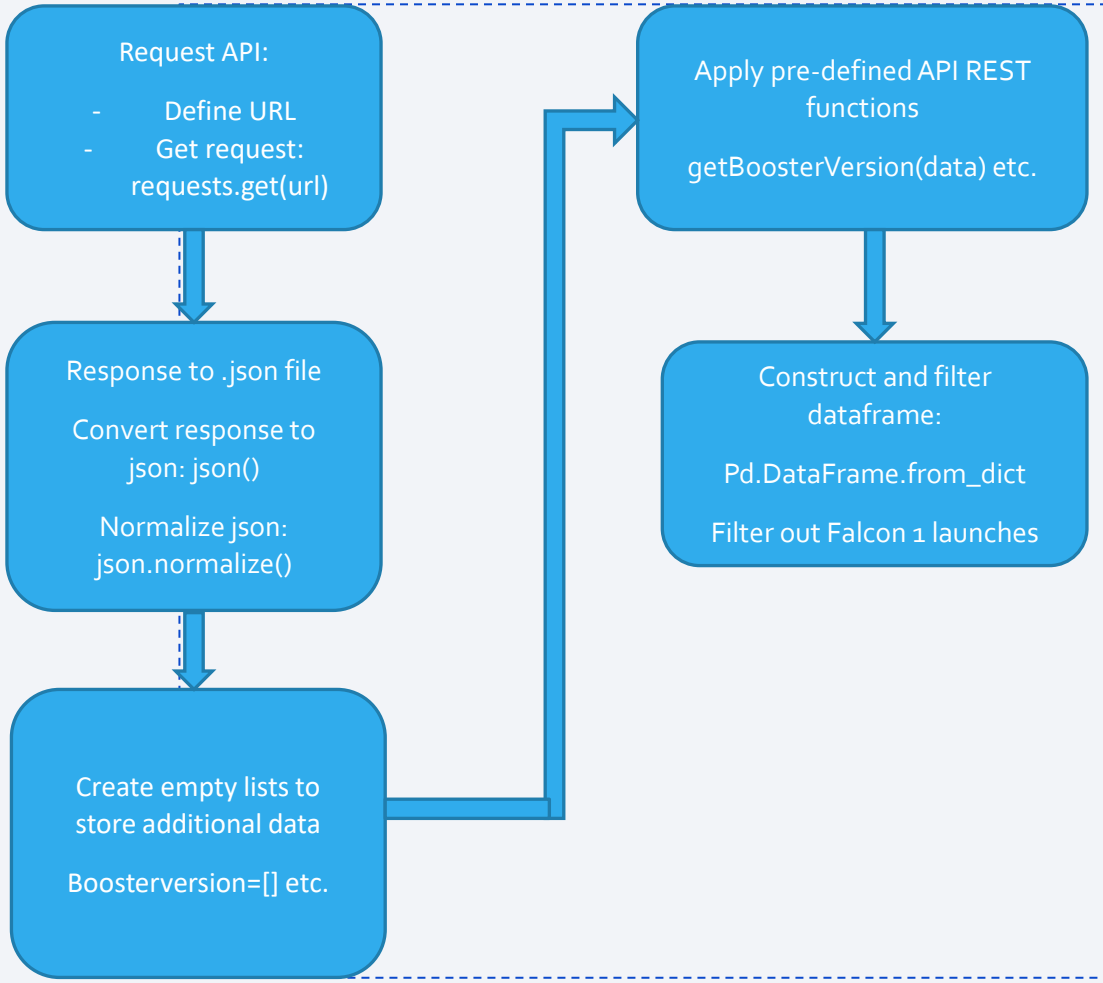
- SpaceX launch data collected from SpaceX API
- Falcon 9 launch data collected from Wikipedia (webscraping)
- General data collection process is following:



Data Collection – SpaceX API

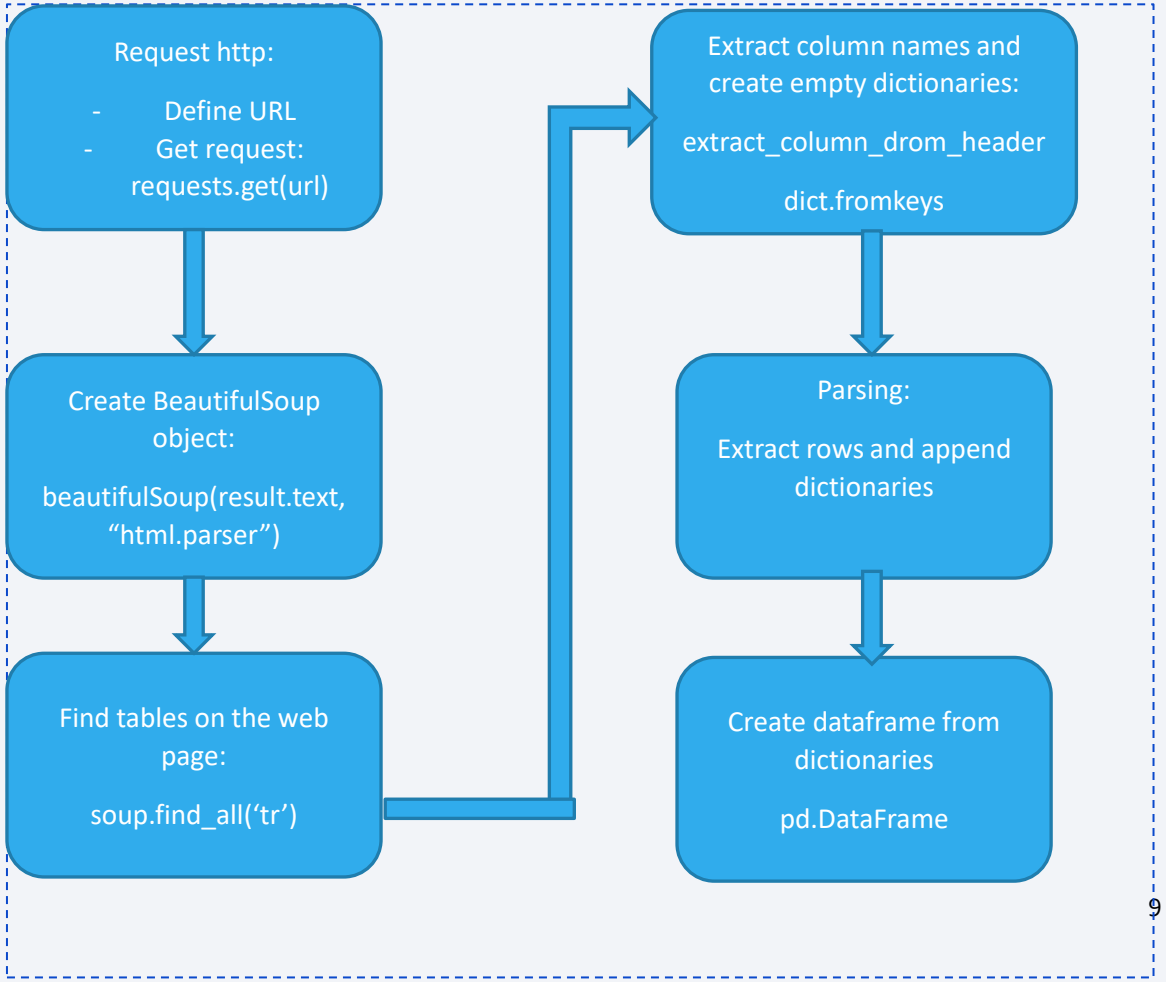
- SpaceXAPI data collection process is indicated on the flowchart

- https://github.com/jaanroos/capstone/blob/main/jupyter-labs-spacex-data-collection-api%20JR.ipynb



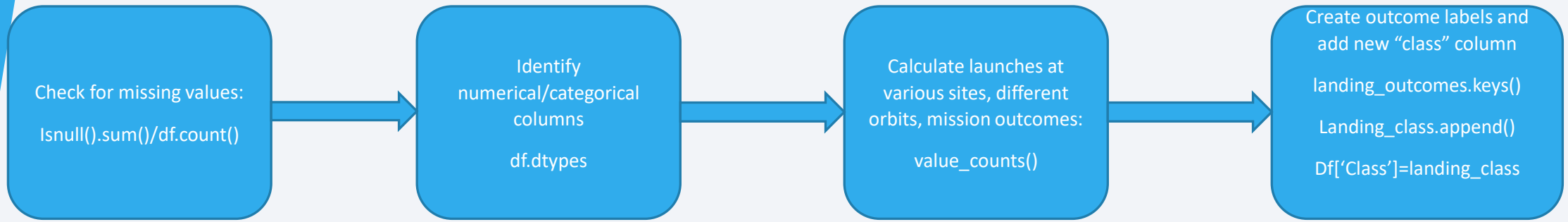
Data Collection - Scraping

- Web data collection process is indicated on the flowchart
- https://github.com/jaanroos/capstone/blob/main/jupyter-labs-webscraping%20JR.ipynb



Data Wrangling

- The main goal of the process is to create landing outcome attribute
- https://github.com/jaanroos/capstone/blob/main/labs-jupyter-spacex-Data%20wrangling_JR.ipynb



EDA with Data Visualization

- Landing result vs flight number and payload mass – identify if payload mass affects landing outcome
- Landing result vs flight number and launch site – identify if launch site affects landing outcome
- Landing result vs payload mass and launch site – identify if there is a relationship between launch site and payload mass
- Landing result vs orbit – identify if orbit type affects landing outcome
- Landing result vs flight number and orbit type – elaborate on the previous graph to see relationship between orbit and landing result
- Landing result vs orbit and payload mass – identify the relationship between orbit and payload mass
- Success rate yearly trend – identify the time factor in landing results

- https://github.com/jaanroos/capstone/blob/main/jupyter-labs-eda-dataviz_JR.ipynb

EDA with SQL

- SQL queries performed:
  - Names of unique launch sites
  - 5 records of launch site beginning with 'CCA'
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - The date of first successful landing outcome in ground pad
  - 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 records of certain columns for the year of 2015
  - The count of successful landings during an indicated period, descending order
- https://github.com/jaanroos/capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite_JR.ipynb

Build an Interactive Map with Folium

- Objects added to the map:
  - Circle and marker of the NASA Johnson Space Centre – testing functionalities
  - Circle and markers of launch site – find launch sites on the map, mark them
  - Marker clusters with color coded markers at launch sites – to visually indicate the success of launches at a specific site
  - Lines and markers (distance in km) from a launch site to proximities – coast, railway, highway, city. Indicates factors to be considered when choosing launch site locations
- https://github.com/jaanroos/capstone/blob/main/lab_jupyter_launch_site_location_JR%20(1).ipynb

Build a Dashboard with Plotly Dash

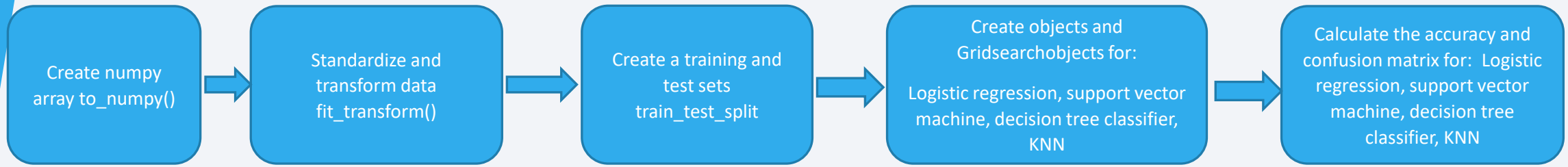
- Contents of the dashboard:
  - Dropdown menu – choose one or all launchsites. The entry point to the dashboard, enables to use one dashboard to analyze every launch site or all of them.
  - Payload mass slider – enables to filter a range of payload mass thus enabling analysis of launches and launch sites by payload mass on a scatterplot
  - Pie chart – indicates the launch rate success at a certain site or share of successful launches for all sites
  - Scatterplot – visualizes success of launches by booster version and payload mass.
- https://github.com/jaanroos/capstone/blob/main/spacex_dash_app_JR_notebook.ipynb

Predictive Analysis (Classification)

- https://github.com/jaanroos/capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5_JR.jupyterlite.ipynb

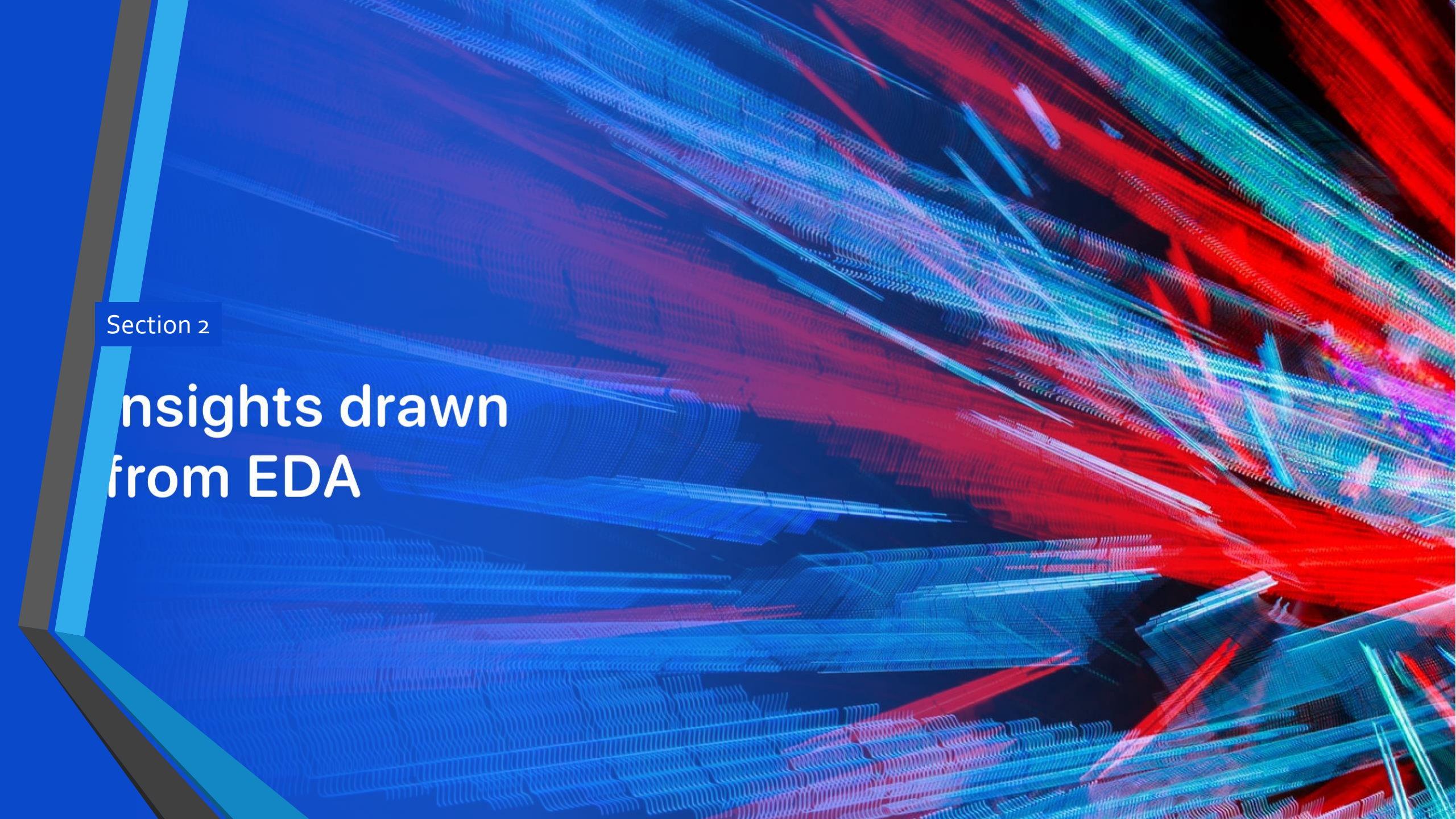
Predictive Analysis (Classification)

- The main goal of the process is to create a predictive model. The process is shown on the flowchart
- https://github.com/jaanroos/capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5_JR.jupyterlite.ipynb



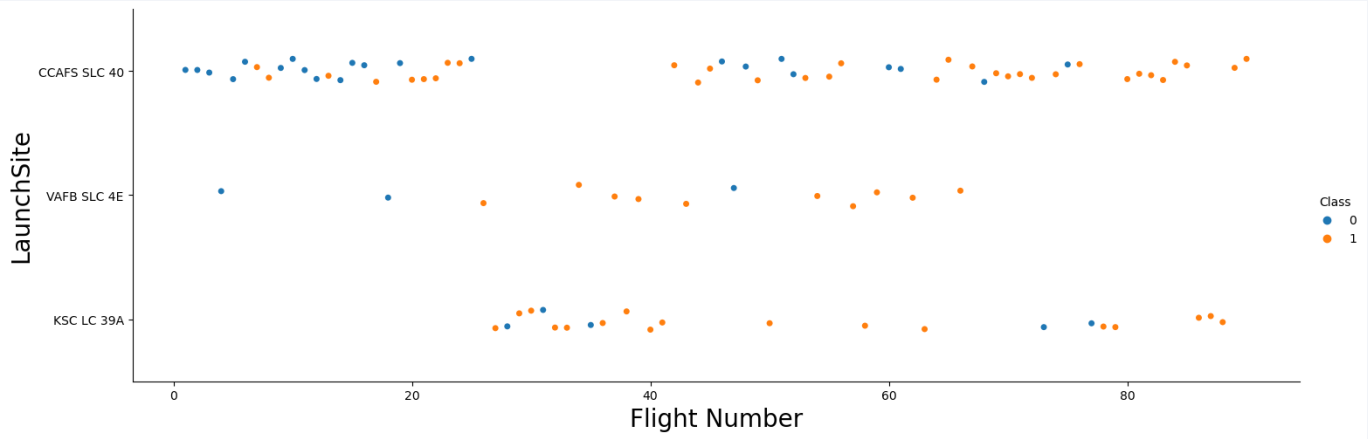
Results

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


Section 2
nsights drawn
from EDA

Flight Number vs. Launch Site

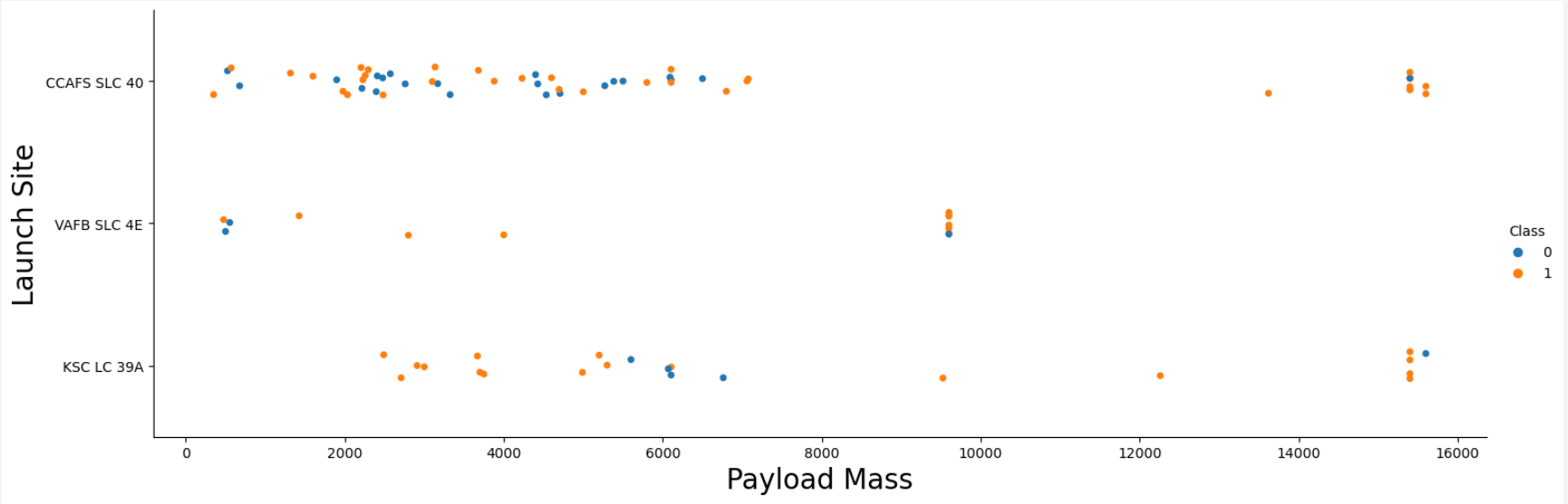
- Show a scatter plot of Flight Number vs. Launch Site



- Two sites have a better success rate
- Later launches (flight number) have a higher success rate

Payload vs. Launch Site

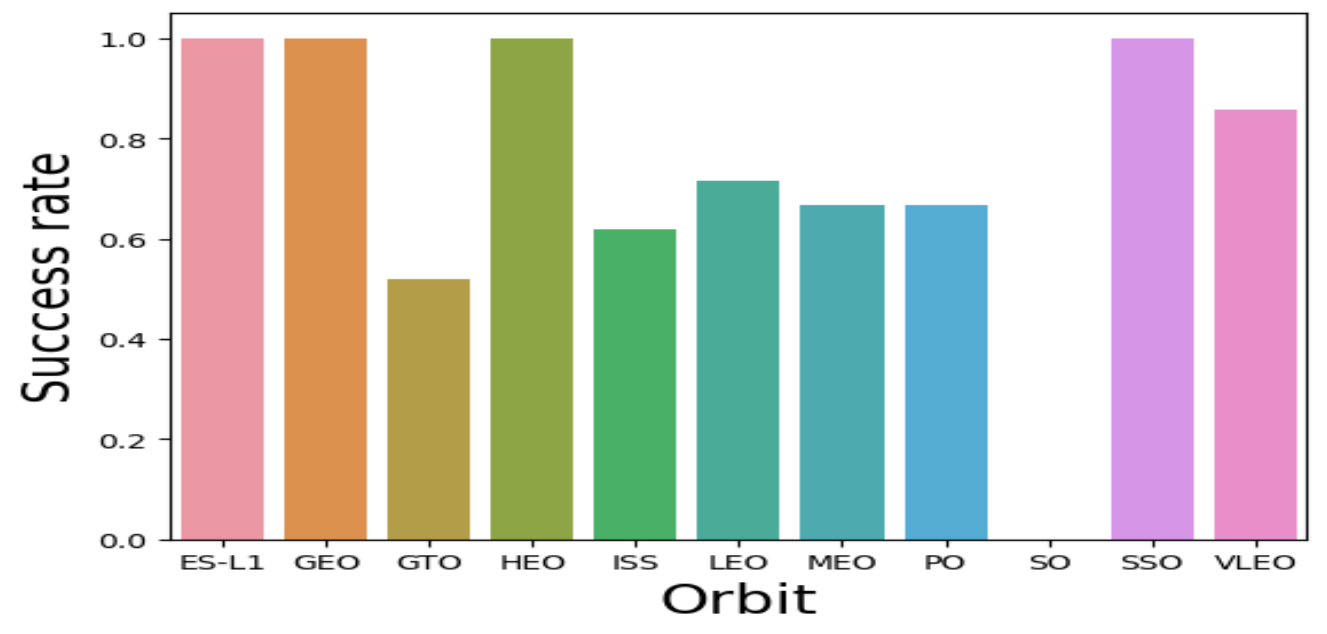
- Show a scatter plot of Payload vs. Launch Site



- KSC LC 39 A has a higher general success rate
- For larger payloads, success rates are similar
- No big payload launches at VAFB SLC 4E

Success Rate vs. Orbit Type

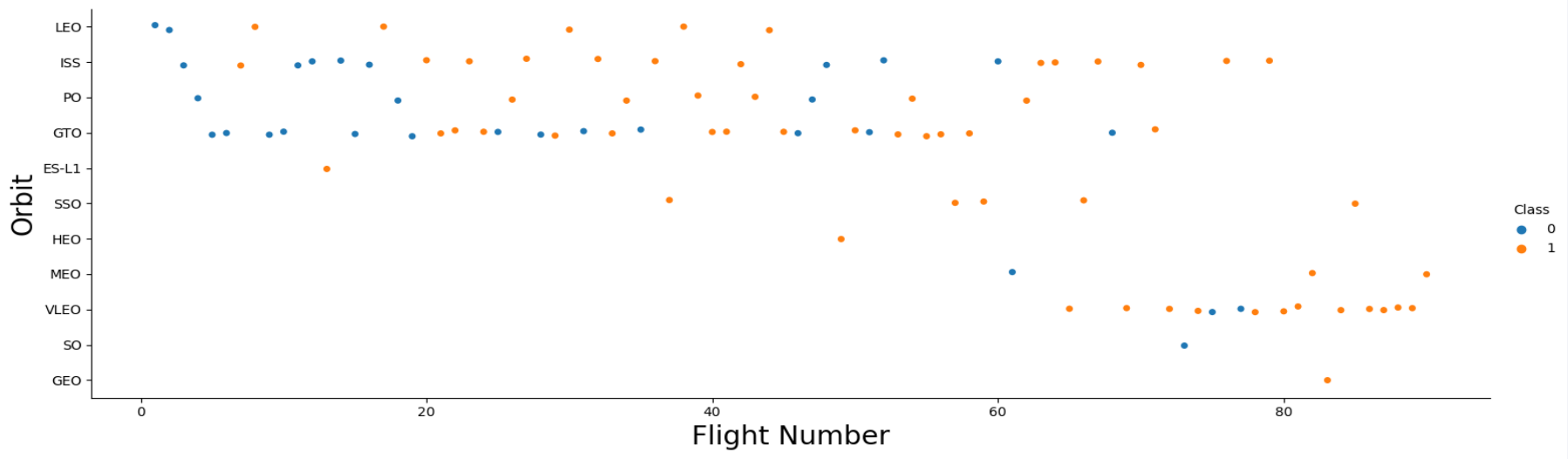
- Show a bar chart for the success rate of each orbit type



- ES-L1, GEO, HEO and SSO have the highest success rates
- GTO has the lowest success rate

Flight Number vs. Orbit Type

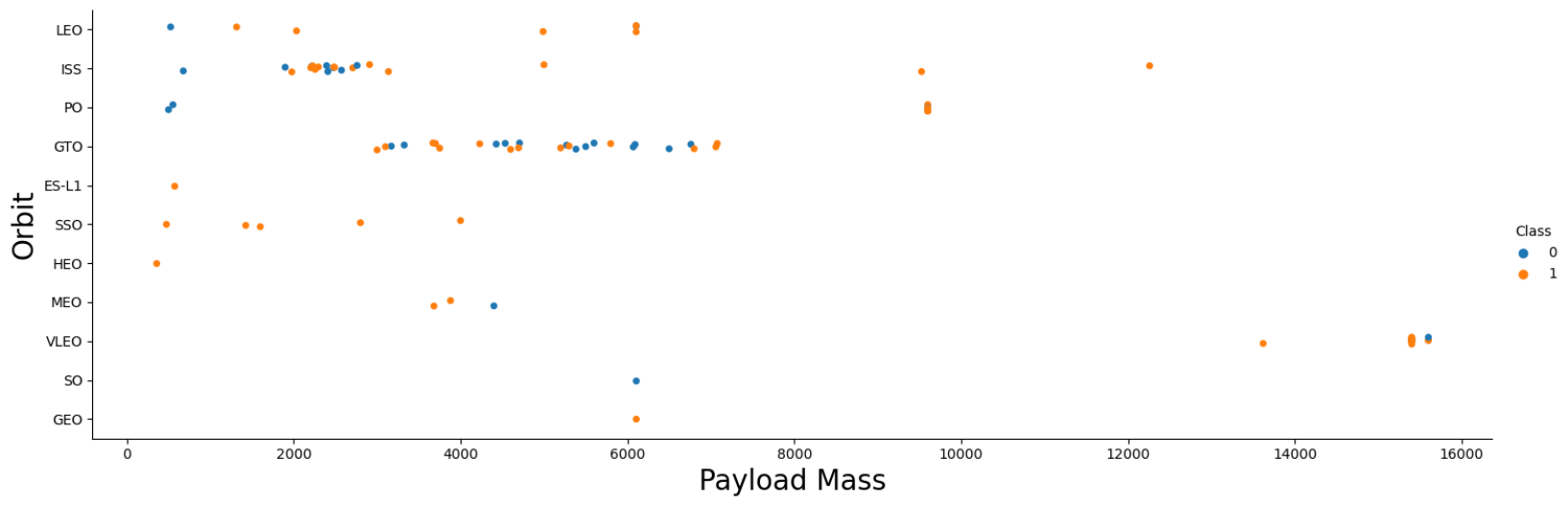
- Show a scatter point of Flight number vs. Orbit type



- GTO has clearly the lowest success rate and has been discontinued
- From earlier flights, ISS, PO and LEO were more successful
- SSO has a perfect success rate but a low number of flights
- VLEO has a good success rate among recent flights and is most commonly used now

Payload vs. Orbit Type

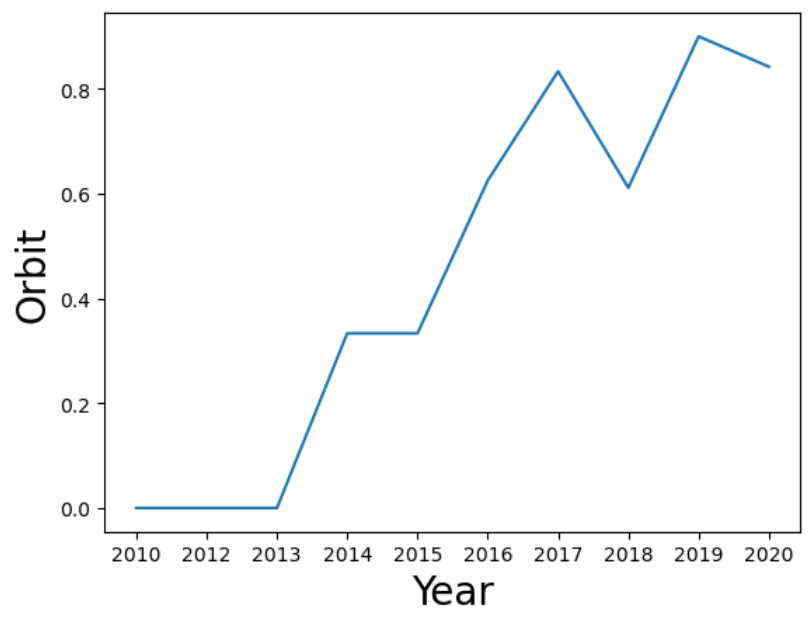
- Show a scatter point of payload vs. orbit type



- VLEO is used for high payload mass launches
- High payload launches have been more successful than light payload launches
- GTO and ISS were mostly used for light payload launches
- SSO and LEO are most successful orbits for light payload

Launch Success Yearly Trend

- Show a line chart of yearly average success rate



- The success rate is showing a constant rise with the exception of 2018

All Launch Site Names

- Four launch sites were identified, some data has no launch site value. DISTINCT clause used.
- %sql SELECT DISTINCT Launch_Site FROM SPACEXTBL

Launch_Site

CCAFS LC-40

VAFB SLC-4E

KSC LC-39A

CCAFS SLC-40

None

Launch Site Names Begin with 'CCA'

- %sql SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5
- WHERE, LIKE and LIMIT clauses were used
- First five results of the query are shown

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Customer LIKE 'NASA (CRS)'
- SUM and WHERE clauses used.
- The result of the query is shown below. It sums the payload mass of NASA booster launches

SUM(PAYLOAD_MASS__KG_)

45596.0

Average Payload Mass by F9 v1.1

- %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version LIKE 'F9 v1.1’
- AVG and WHERE clauses were used.
- Result below shows the average payload of all launches

AVG(PAYLOAD_MASS__KG_)

2928.4

First Successful Ground Landing Date

- %sql SELECT MIN(Date) FROM SPACEXTBL WHERE Landing_Outcome LIKE 'Success (ground pad)'
- MIN, WHERE and LIKE clauses were used.
- The query result shows the earliest successful ground pad landing

MIN(Date)

01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

- %sql SELECT DISTINCT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome LIKE 'Success (drone ship)' AND (PAYLOAD_MASS__KG_>4000 AND PAYLOAD_MASS__KG_<6000)
- DISTINCT, WHERE, LIKE, AND clauses were used.
- The query results show successful drone ship landings for payload between 4000 and 6000 kg.

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

Total Number of Successful and Failure Mission Outcomes

- %sql SELECT Mission_outcome, COUNT (*) FROM SPACEXTBL GROUP BY Mission_Outcome
- COUNT and GROUP BY clauses were used.
- Various mission outcomes are shown below:

Mission_Outcome	COUNT (*)
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- %sql SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
- WHERE, MAX clauses and a sub-query were used.
- The query result is shown below:

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

2015 Launch Records

- %sql SELECT substr(Date, 4, 2) as MONTH, Booster_Version, Launch_Site FROM SPACEXTBL WHERE substr(Date, 7, 4)='2015' AND Landing_Outcome LIKE 'Failure (drone ship)’
- WHERE, substr, DATE, AND, LIKE clauses were used
- The query results are shown below:

MONTH	Booster_Version	Launch_Site
10	F9 v1.1 B1012	CCAFS LC-40
04	F9 v1.1 B1015	CCAFS LC-40

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

- %sql SELECT Landing_Outcome, COUNT(Landing_Outcome) FROM SPACEXTBL WHERE Landing_Outcome LIKE '%Success%' AND substr(Date,7,4)||substr(Date,4,2)||substr(Date,1,2) between '20100406' AND '20170320' GROUP BY Landing_Outcome ORDER BY COUNT(Landing_Outcome) DESC
- COUNT, WHERE, AND, substr, DATE, BETWEEN, AND, GROUP BY, ORDER BY and DESC clauses were used.
- The results are:

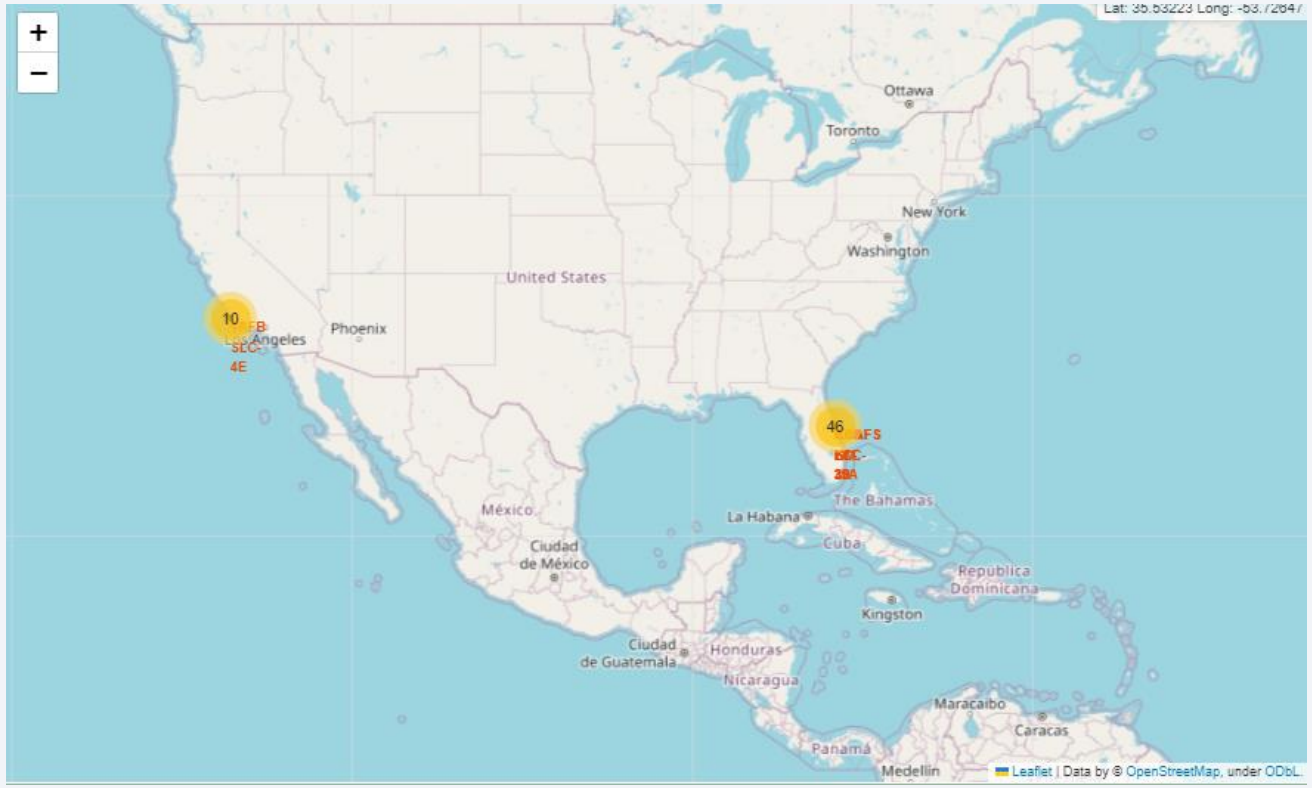
Landing_Outcome	COUNT(Landing_Outcome)
Success (ground pad)	5
Success (drone ship)	5

Section 3

Launch Sites Proximities Analysis

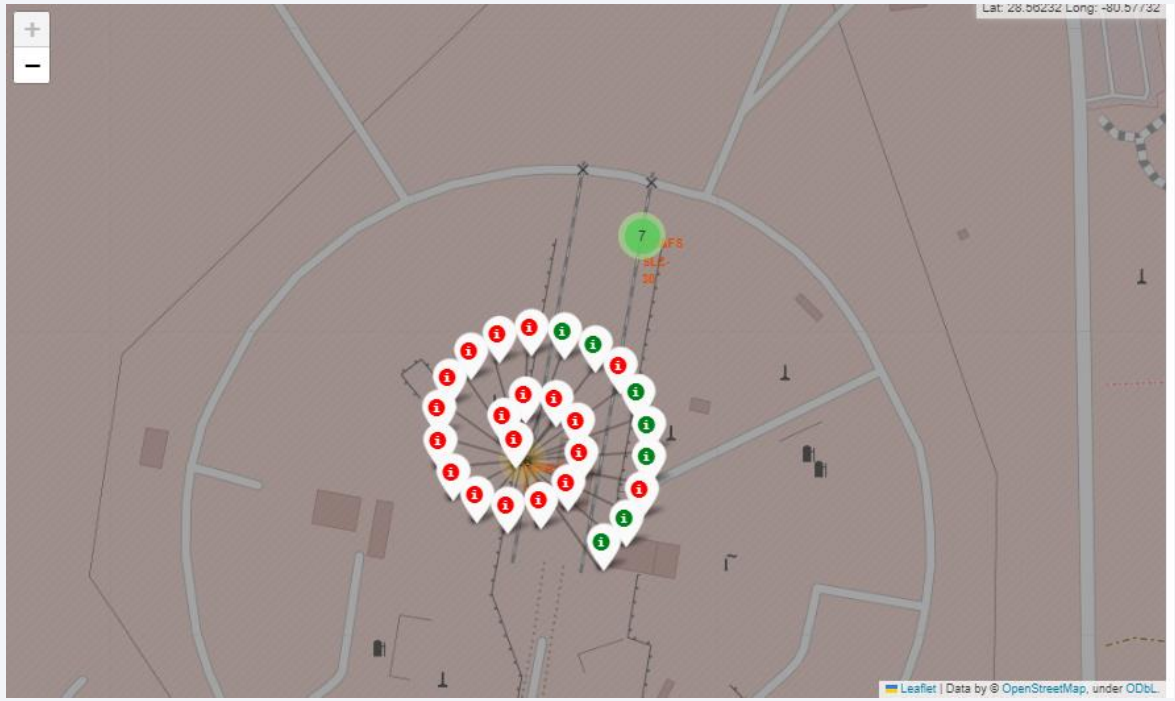
Launch site locations

- Launch sites are located close to oceans and in the southern part of the US



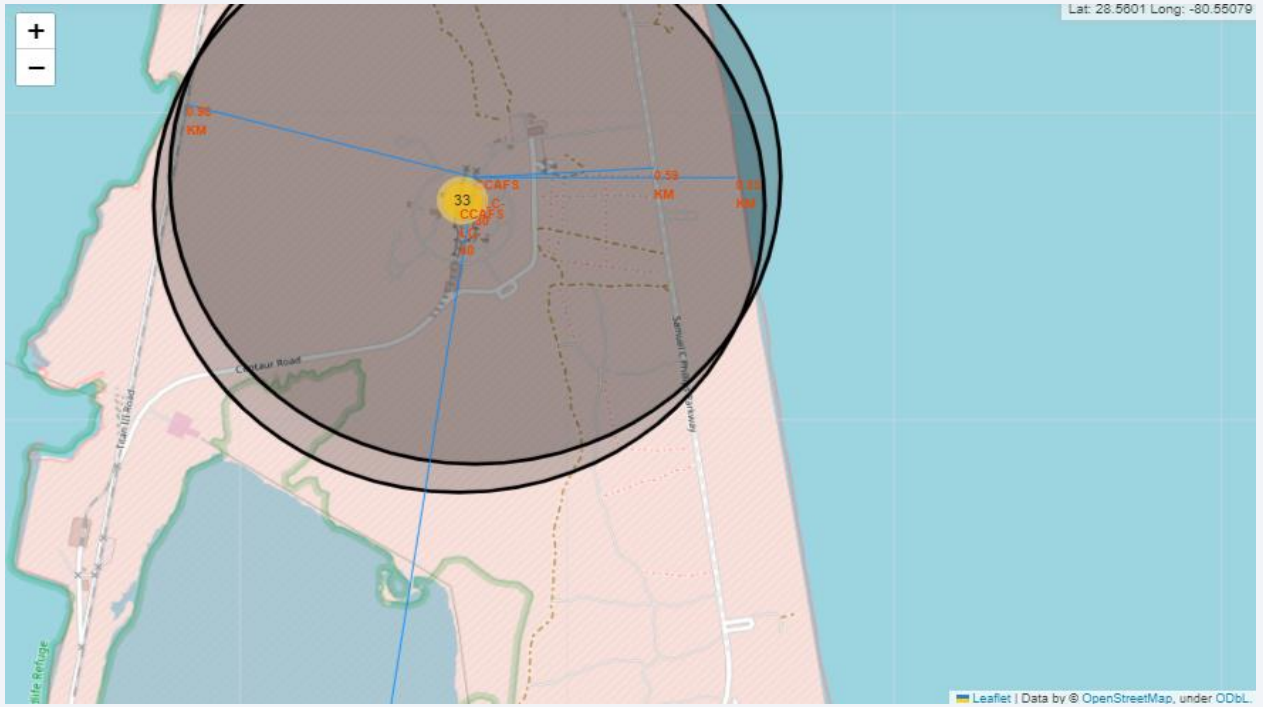
Launch success markers

- Launch success markers are added to all launch sites. KSC LC 39A has the best success rate. An example is given below:



Proximity area of a launch site

- Markers (distance) and lines were added to indicate distance to objects from a launch site
- Launch site is in proximity of an ocean, highway and railway but some distance from the nearest city.

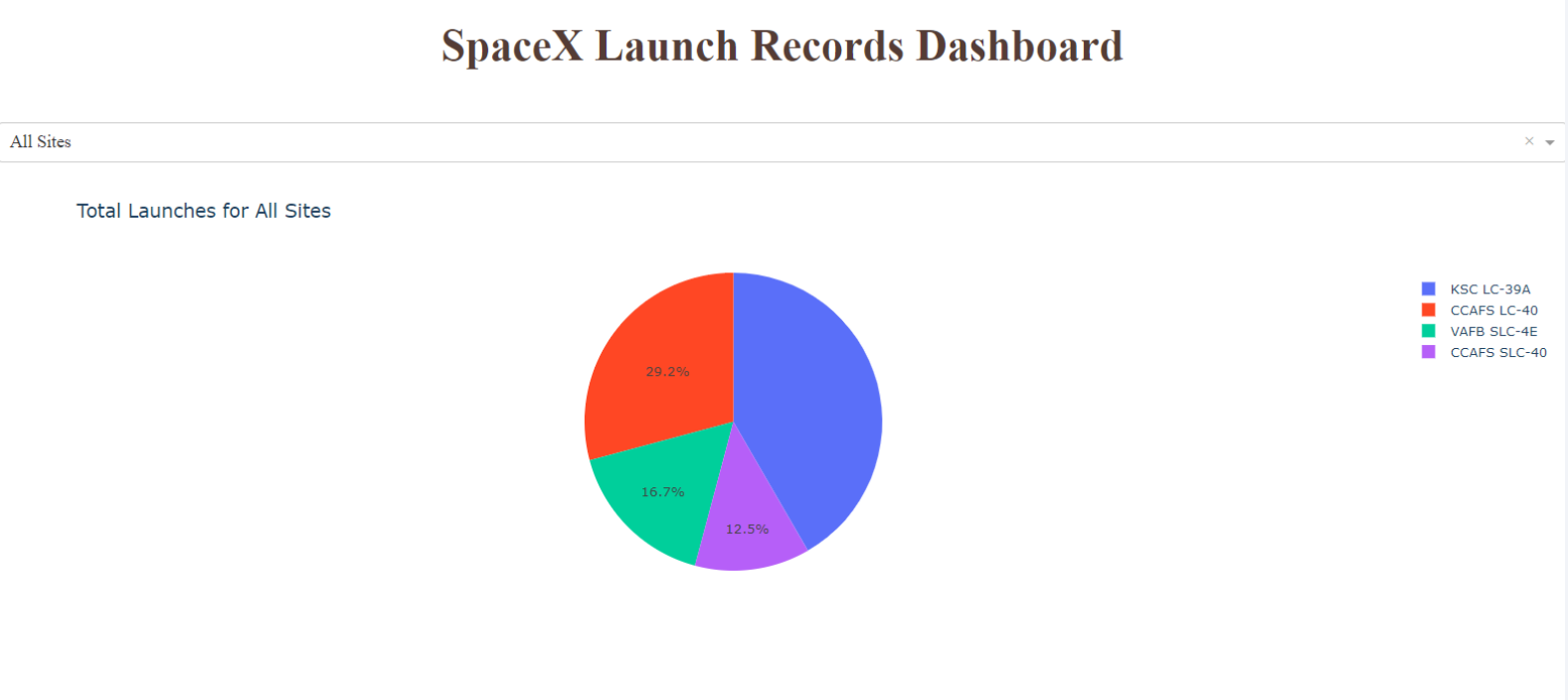


Section 4

Build a Dashboard with Plotly Dash

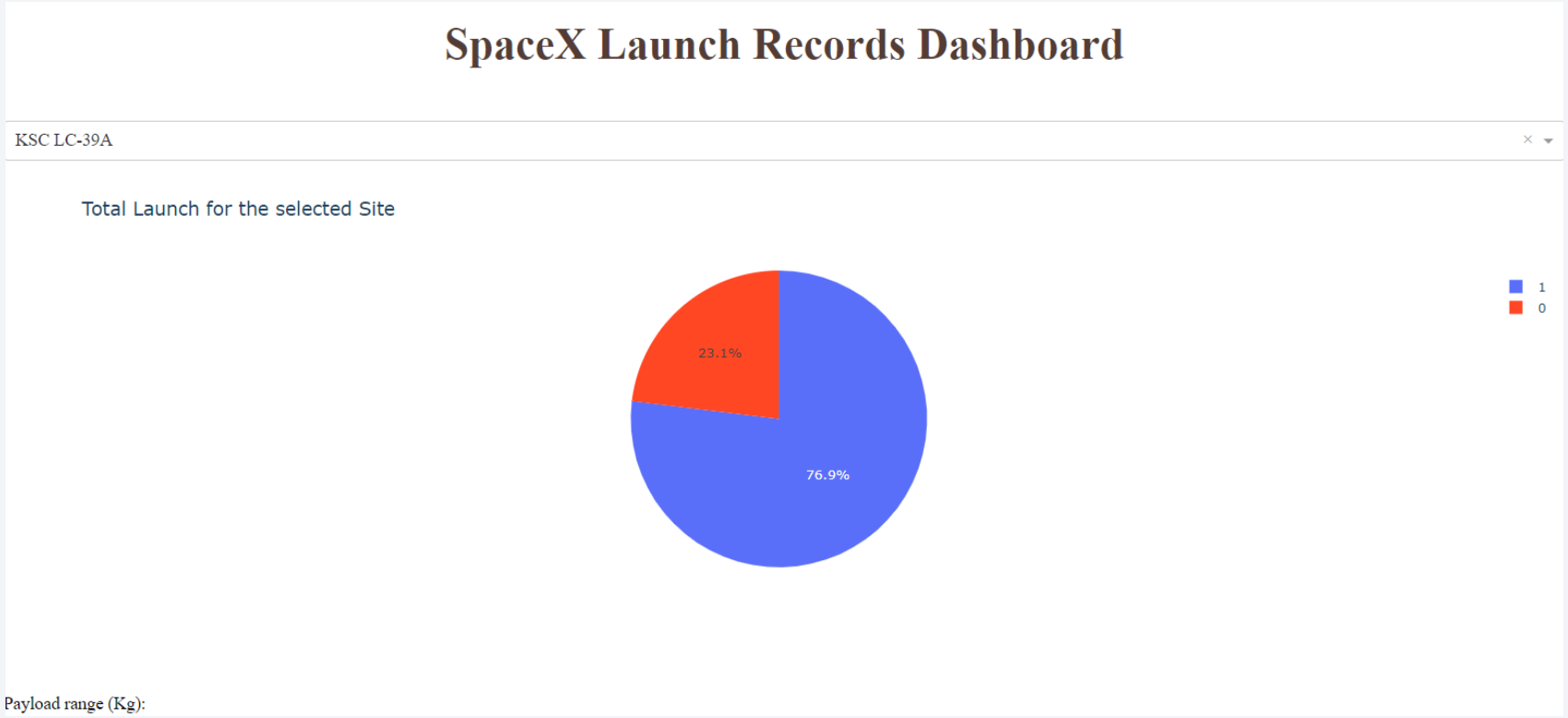
Share of successful launches by launch site

- KSC LC-39A had the most successful launches



Launch site with the highest success rate

- KSC LC-39A also has the highest success rate among launch sites



Low vs high payload booster success

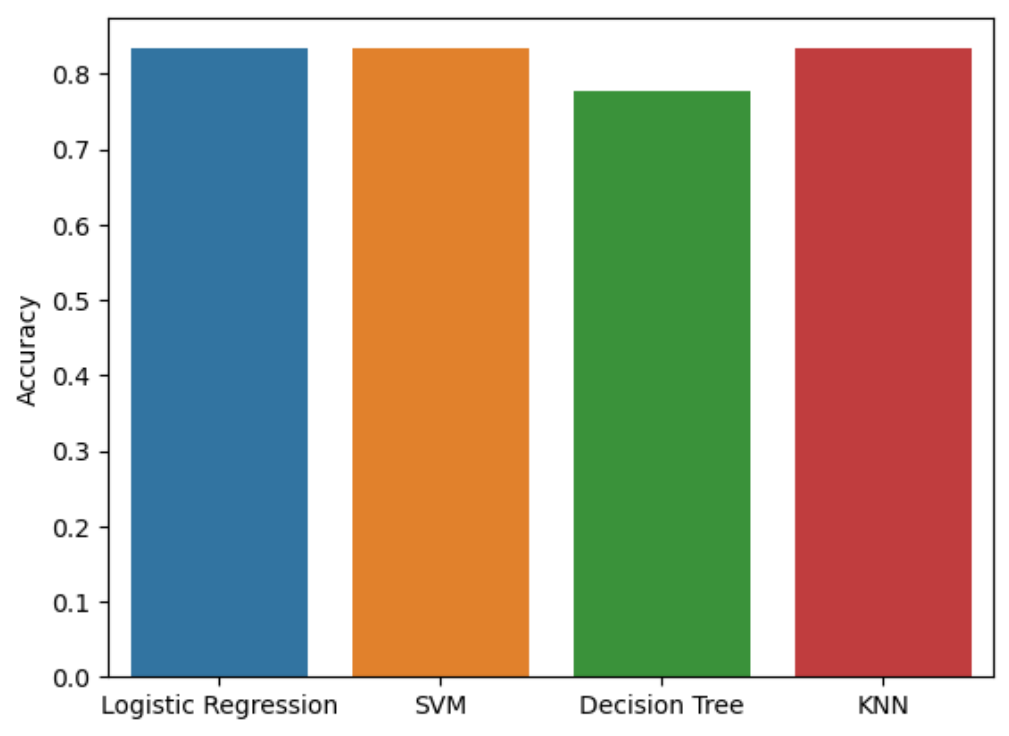
- There are very few launches with heavier payload – with a low rate of success
- For lighter payload FT and B4 boosters show a better success rate.

Section 5

Predictive Analysis (Classification)

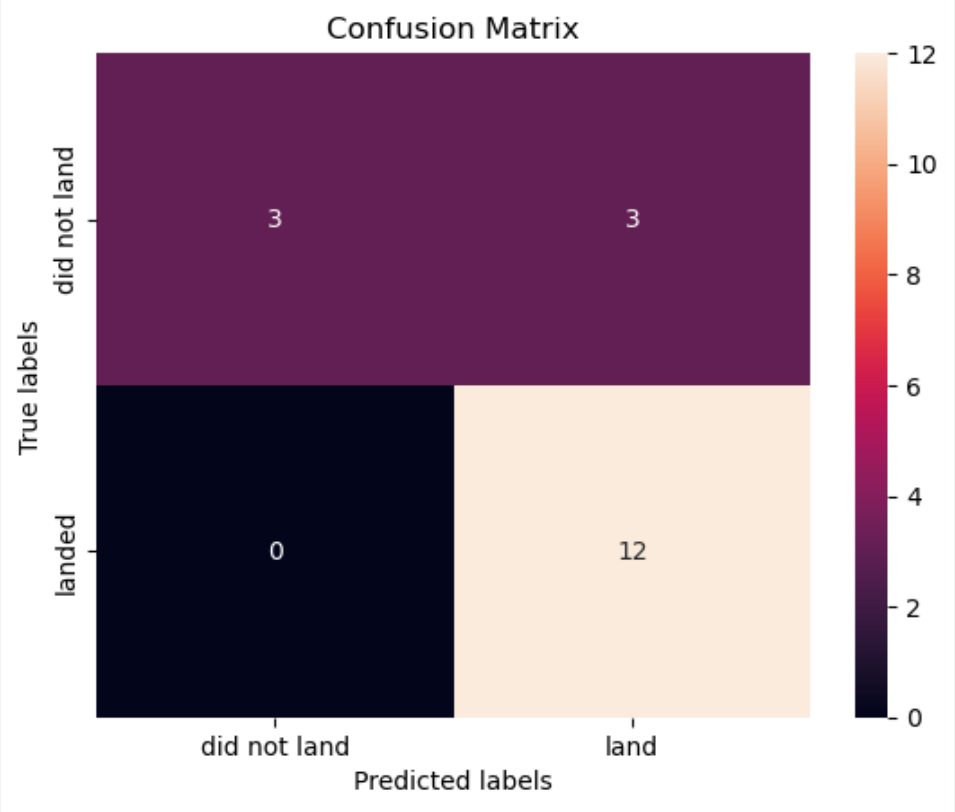
Classification Accuracy

- Decision tree shows lowest accuracy on the test data. Other methods are equal.



Confusion Matrix

- SVM, KNN and logistic regression produce a similar confusion matrix. While precise in predicting landings, the problem remains false positives i.e. predicting successful landing whereas failure happens.



Conclusions

- The success of SpaceX launch is directly related to the year it was performed
- Light payloads perform better than heavy payloads
- SVM, KNN and Logistic Regression models are the best predictive models for this dataset
- KSC LC 39 A has the most successful launches and the best launch success ratio
- Orbits ES L1, GEO, HEO, SSO, VLEO have best success rates.

Appendix

- Github repository link with materials:

https://github.com/jaanroos/capstone

Thank you!