Winning Space Race with Data Science

Iury Benevelli
2023-07-01

Outline

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

Executive Summary

- Data analysis methodologies
  - Data collection (SpaceX API and Wikipedia web scraping);
  - Exploratory data analysis (data wrangling, data visualization ad interactive visuals)
  - Machine learning predictions
- Summary of all results
  - Data collection from public sources
  - EDA allowed to identify the best features to predict success
  - Machine learning predictions showed the best model to use

Introduction

- The objective is to evaluate the viability of the new company Space Y to compete with Space X.
- Problems you want to find answers
  - Best total cost estimation for launches
  - Best place for launches

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from SpaceX was obtained from 2 sources
    - SpaceX API (https://api.spacexdata.com/v4/rockets/)
    - Webscraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_Launches)
- Perform data wrangling
  - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features
- Perform exploratory data analysis (EDA) using visualization and SQL

Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data that was collected until this step were normalized, divided in training and test data sets and evaluated by four different classification models, being the accuracy of each model evaluated using different combinations of parameters.

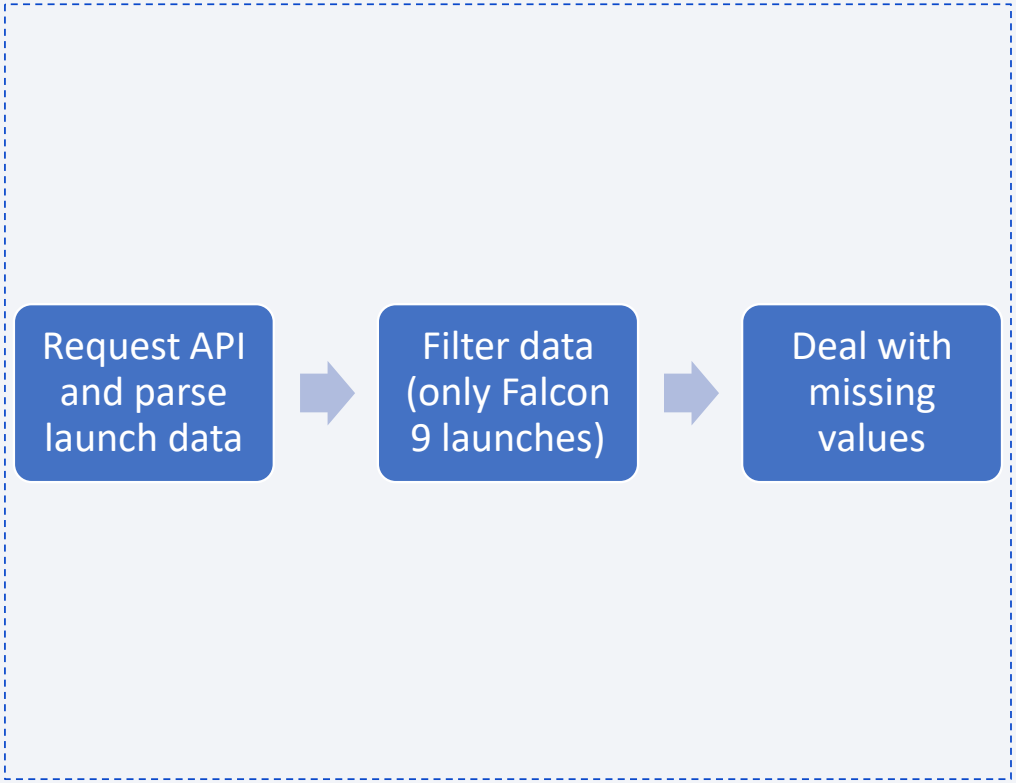
Data Collection

- Data sets were collected from Space X API (https://api.spacexdata.com/v4/rockets/rockets/) and from Wikipedia https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using web scraping technics.

Data Collection – SpaceX API

- Data obtained from a public API offered by SpaceX

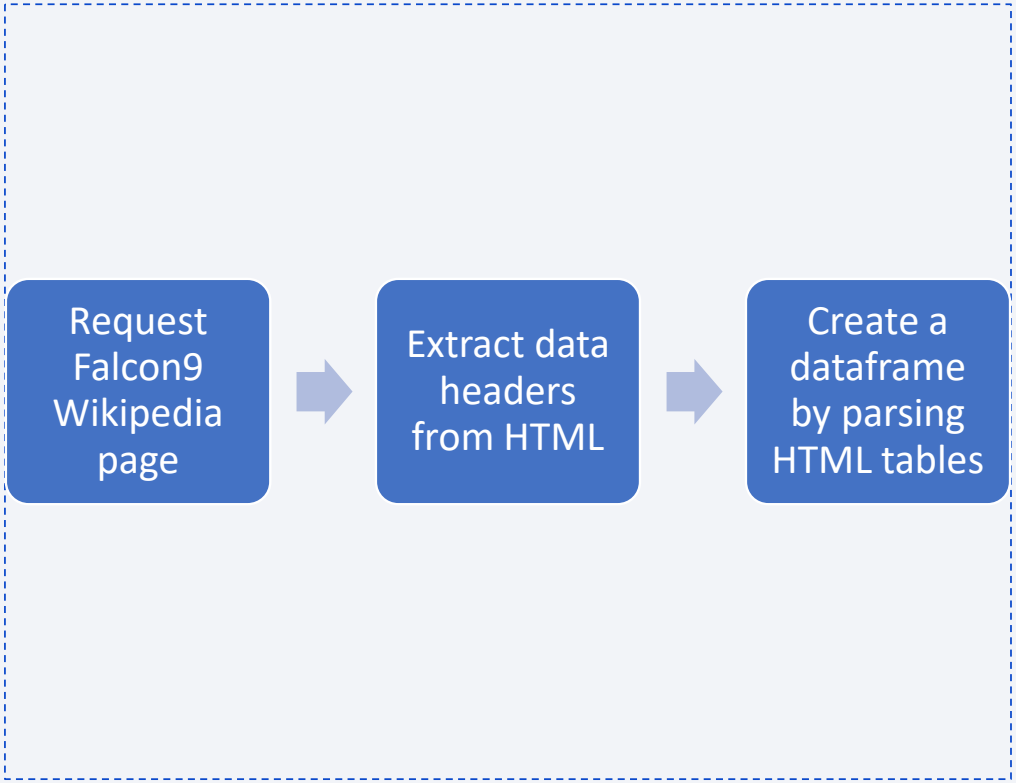
- Source code: https://github.com/benva85/IBM_Datascience_Capstone/blob/main/01_spacex-data-collection-api.ipynb



Data Collection - Scraping

- Data can also be obtained from Wikipedia via web scraping

- Source code: https://github.com/benva85/IBM_Datascience_Capstone/blob/main/02_webscraping.ipynb



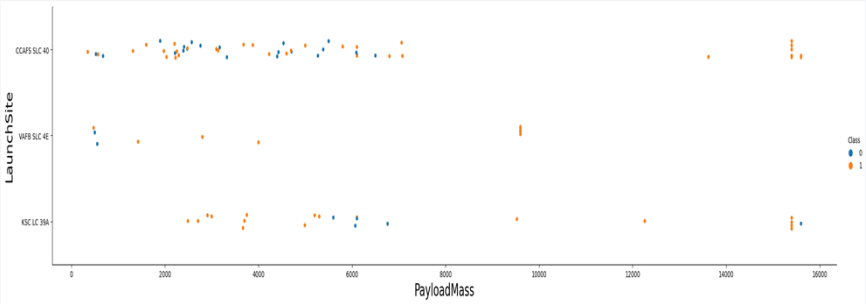
Data Wrangling

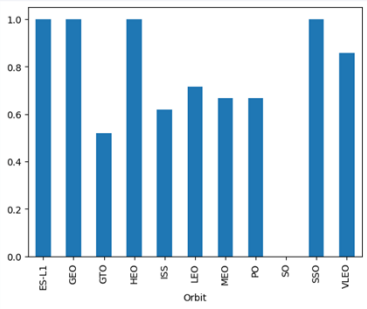
- Initial EDA on the dataset
- Data summarization
  - Launches per site
  - Orbit type
  - Mission outcome per orbit type
- Created landing outcome label

- Source code: https://github.com/benva85/IBM_Datascience_Capstone/blob/main/03_spacex-data_wrangling_.ipynb

EDA with Data Visualization

- Scatterplots and barplots were used to explore data, visualizing relationships between features
  - Payload mass_vs_flight number, Launch site_vs_flight number, …, …





- Source code: https://github.com/benva85/IBM_Datascience_Capstone/blob/main/05_eda-dataviz.ipynb

EDA with SQL

- Several SQL queries were performed to explore the dataset:
  - Unique launch sites names in the space mission
  - Top 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
  - date when the first succesful landing outcome in ground pad was achieved
  - names of the 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 the booster_versions which have carried the maximum payload mass
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for inyear 2015
  - 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
- Source code: https://github.com/benva85/IBM_Datascience_Capstone/blob/main/04_eda-sql.ipynb

Build an Interactive Map with Folium

- Folium maps were used to display:
  - Markers that indicate launch sites and interest points
  - Circle which indincate important areas around specific coordinates (e.g. NASA JSC)
  - Marker cluster used to group events at a certain coordinate
  - Lines to indicate distances between interest points

- Source code: https://github.com/benva85/IBM_Datascience_Capstone/blob/main/06_launch_site_location.ipynb

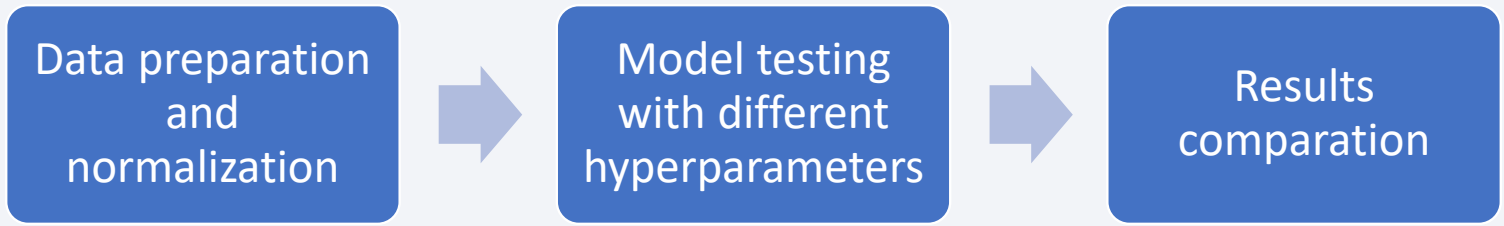
Build a Dashboard with Plotly Dash

- The following plots were used to visualize data
  - Percentage of launches by site
  - Payload range
- This plots allowed to quickly analyze the relationships between launch site and payload, in the mean to identify the best place to launch based on payloads

- Source code: https://github.com/benva85/IBM_Datascience_Capstone/blob/main/07_spacex_dash_app.py

Predictive Analysis (Classification)

- Four classification methods were compared:
  - Logistic regression
  - SVM
  - Decision trees
  - KNN



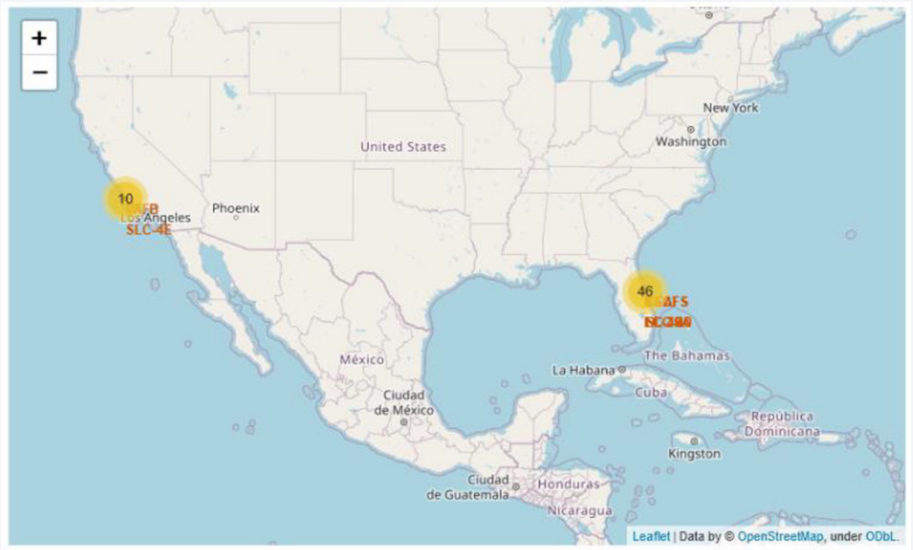
- Source code: https://github.com/benva85/IBM_Datascience_Capstone/blob/main/08_SpaceX_Machine_Learning_Prediction.ipynb

Results

- Exploratory data analysis results
  - 4 different launch sites
  - Average booster payload 2.928 kg
  - First successful landing happened in 2015, 5 years after the first launch
  - Many booster version obtained successful landings having payloads above average
  - Almost 100% success rate
  - Booster versions v1.1 B1012 and v1.1 B1015 failed to land in 2015
  - Landing success rate incrased as years passed

Results

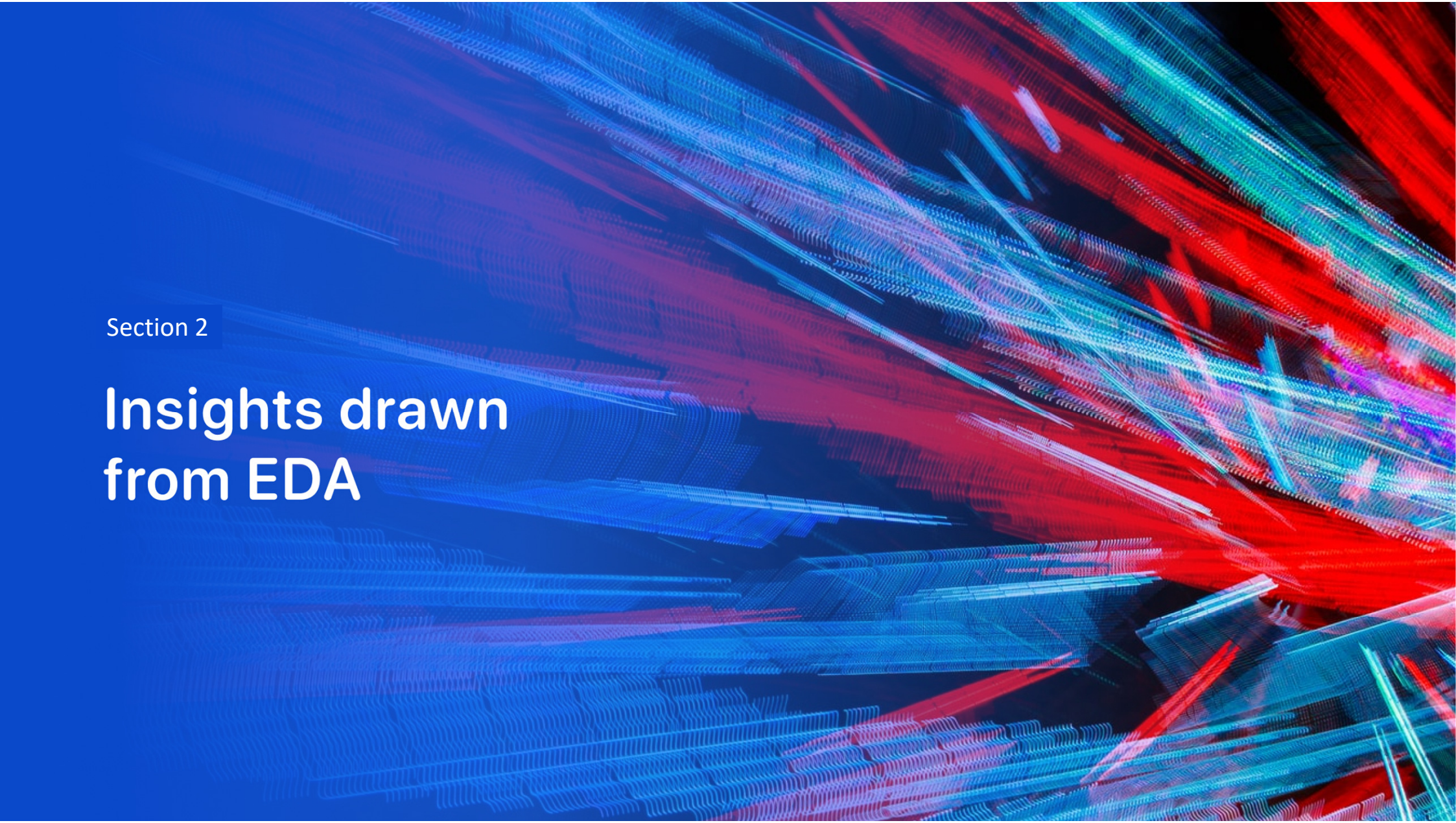
- Using interactive map analytics was possible to identify that launch sites are near the sea and near important infrastructures and facilities; also launch sites are as near to the equator as possible
- East coast launch sites have seen the most launches



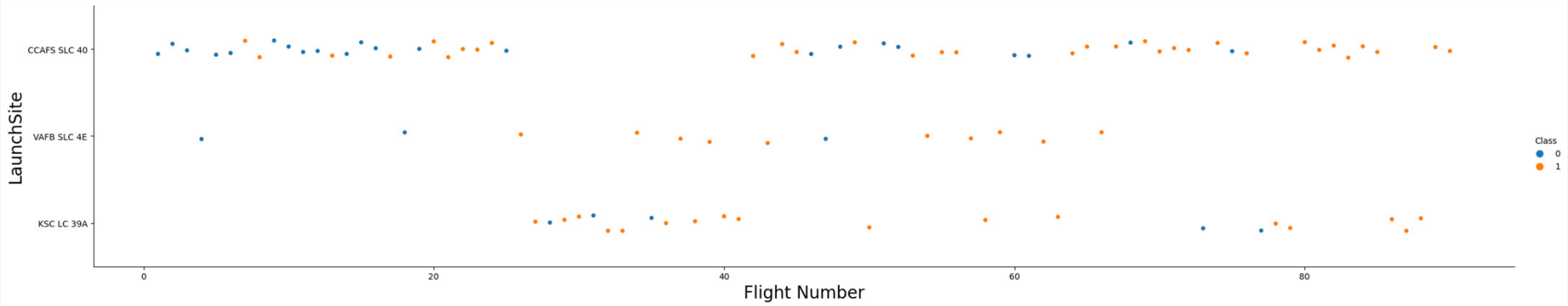
Results

- Predictive analysis showed that decision tree classifier is the best model to predict successful landings

Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.90536	0.88889
KNN	0.84821	0.83333

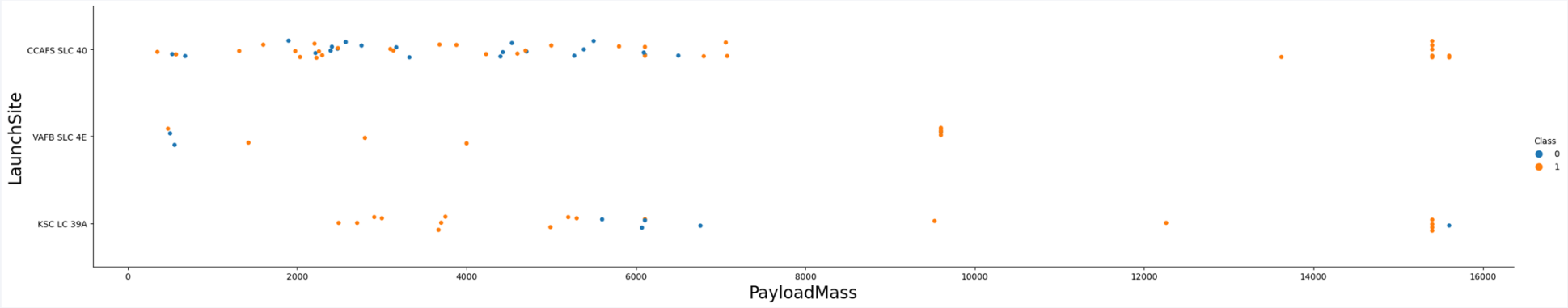

Section 2
Insights drawn
from EDA

Flight Number vs. Launch Site



- CCAFS SLC 40 is the best launch site, most recent launches in this launch site wewe successful
- General success improved over time

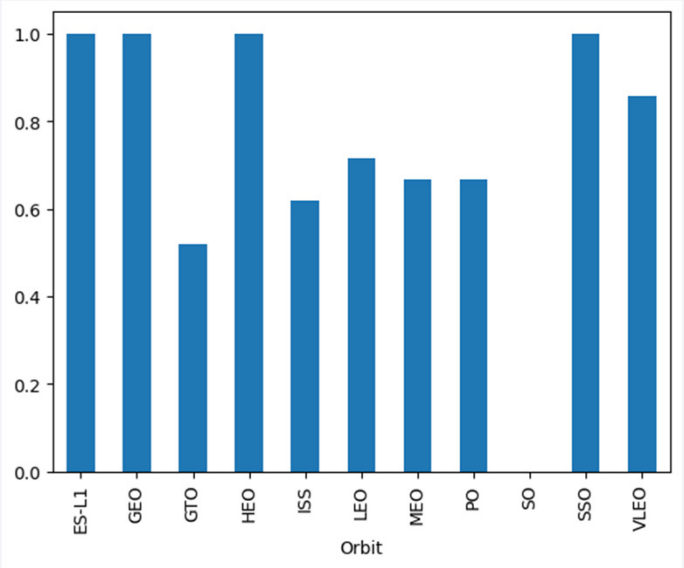
Payload vs. Launch Site



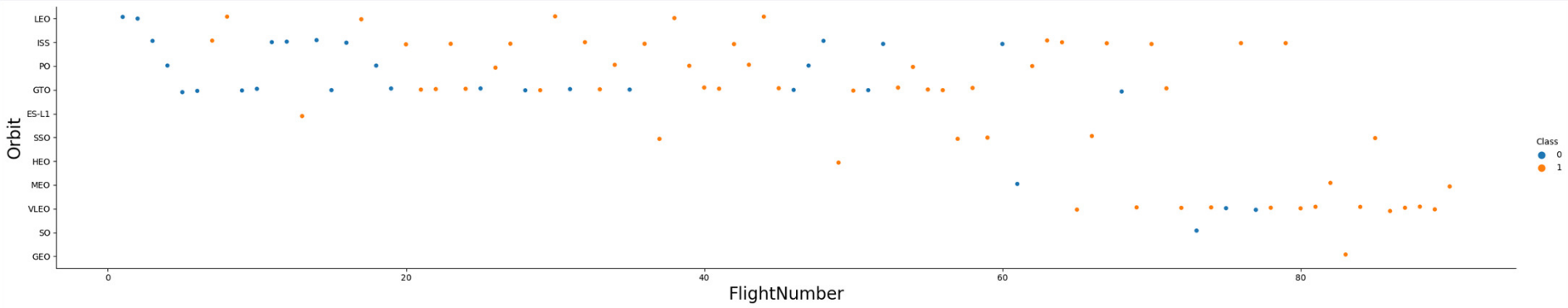
- Payloads over 9.000kg have better success rate
- No launches with payloads >= 10.000 kg were done on VAFB SLC 4E launch site

Success Rate vs. Orbit Type

- Best orbits based on success rate:
  - ES-L1
  - GEO
  - HEO
  - SSO
- VLEO and LEO orbits also have acceptable success rates

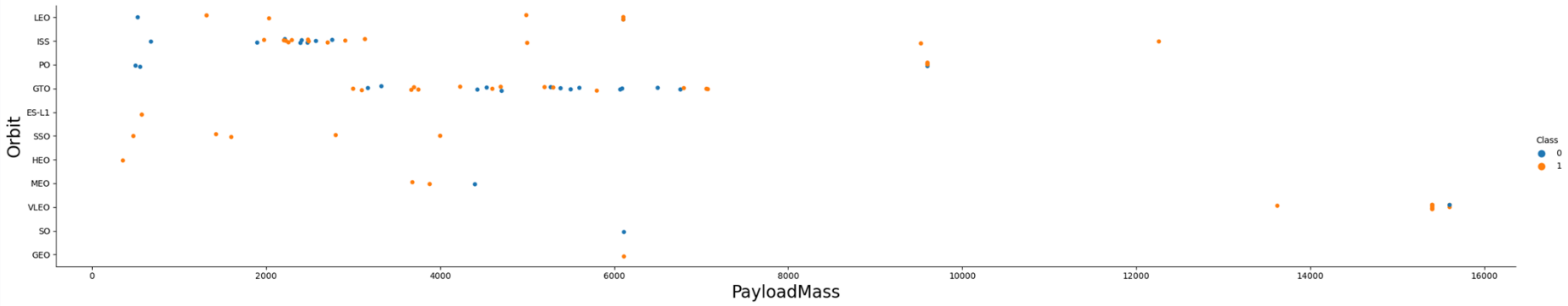


Flight Number vs. Orbit Type



- Success rate impoves over time for all orbit types
- VLEO orbit launch frequency increased in most recent launches

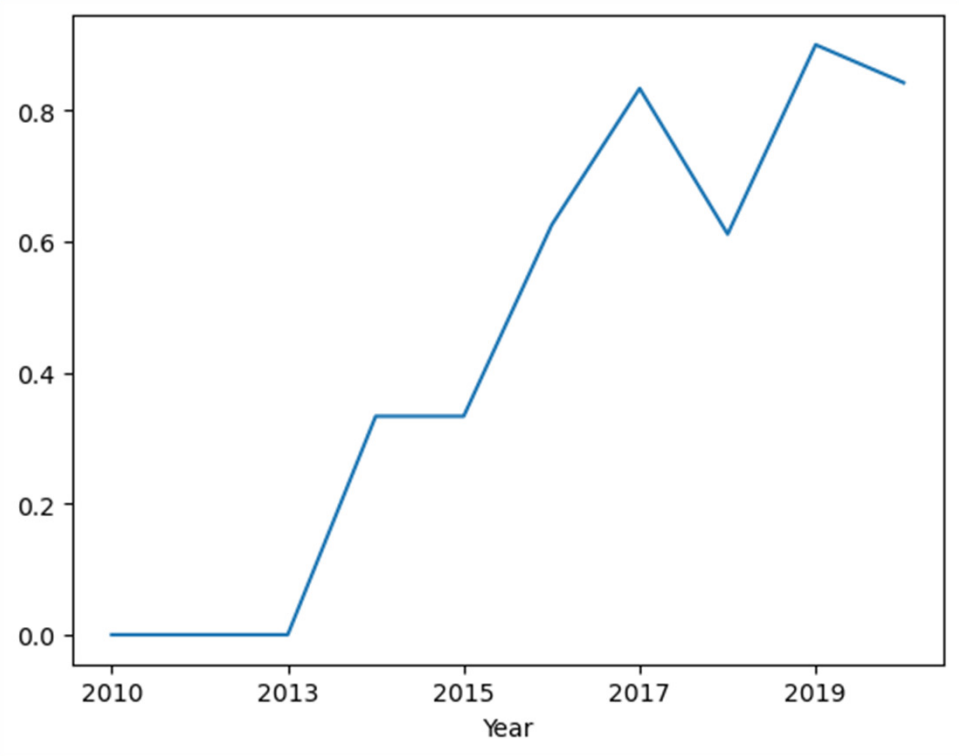
Payload vs. Orbit Type



- SO and GEO orbits are the less used
- SSO orbit always successful on payloads up to 4.000 kg
- GTO and ISS orbits have the widest range of payloads and success rates

Launch Success Yearly Trend

- Success rate started to increase steadily from 2013 to 2020
- First 3 years of launches were used to develop the technologies and success rate never increased



All Launch Site Names

- 4 launch sites are listed in the dataset

Launch Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39
AVAFB SLC-4E

- Select unique occurences in the “launch_site” values from the dataset

Launch Site Names Begin with 'CCA'

- Find 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
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	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	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Five samples of Cape Canaveral launches

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

TOTAL_PAYLOAD
111268

- Total payload calculated by summing all payloads whose code contain ‘CRS’ corresponding to NASA

Average Payload Mass by F9 v1.1

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

AVG_PAYLOAD
2928.4

- Filtering data by booster version and calculating average payload mass

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

FIRST_SUCCESS_GP
01/08/2018

- By filtering data by successful landing outcome on ground pad and getting the minimum value for date it's possible to identify the first occurrence

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

- Selecting distinct booster versions according to the filters above, these 4 are the result.

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

Mission_Outcome	QTY
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- Grouping mission outcomes and counting records for each group led us to the summary above.

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

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

2015 Launch Records

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

Booster_Version	Launch_Site	month
F9 v1.1 B1012	CCAFS LC-40	10
F9 v1.1 B1015	CCAFS LC-40	4

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

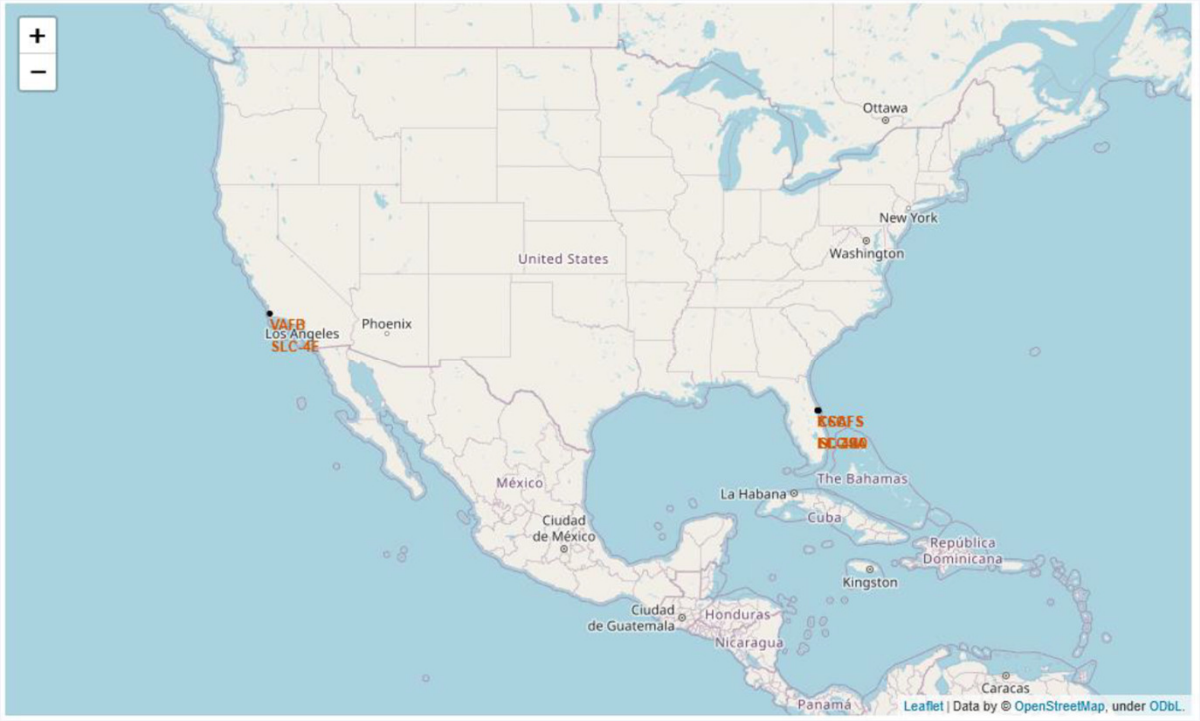
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

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

- A lot of “no attempt”

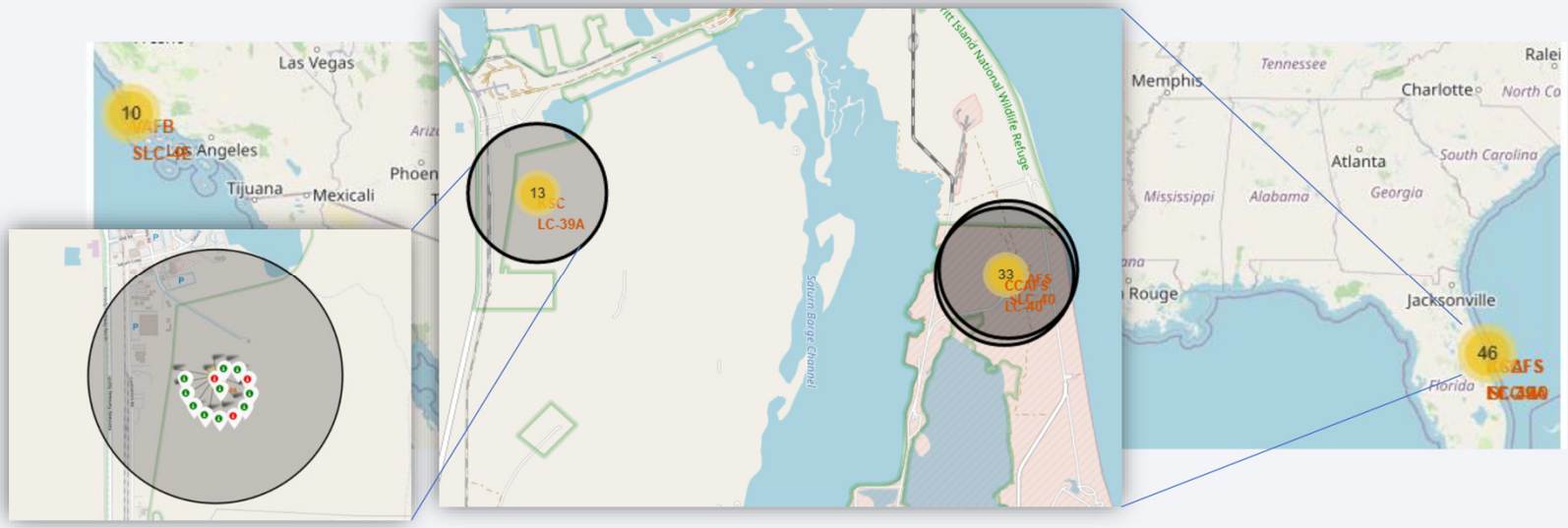

Section 3
Launch Sites
Proximities Analysis

All launch sites



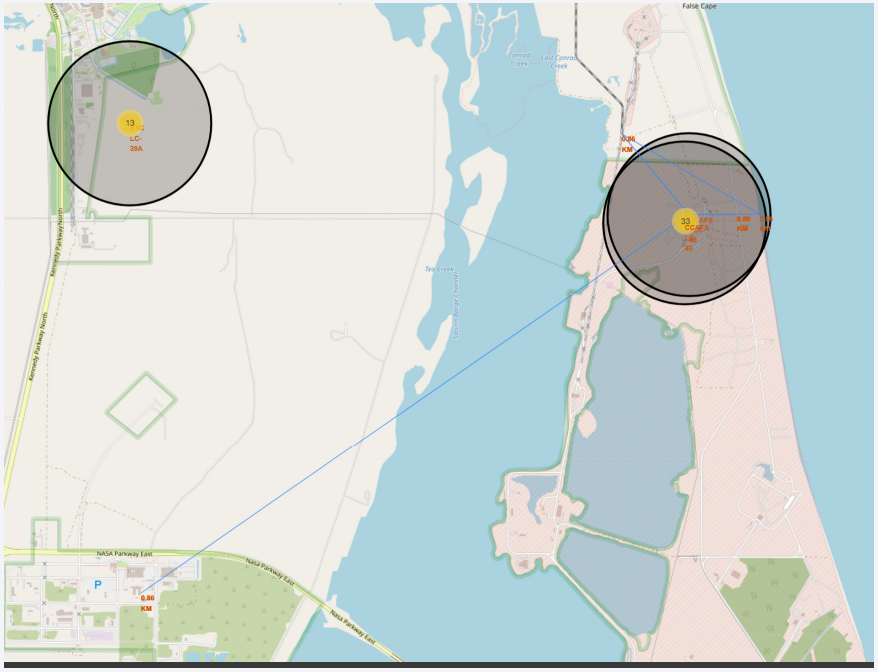
- All launch sites are near sea and near the equator for safety and orbital reasons

Launch outcome by site

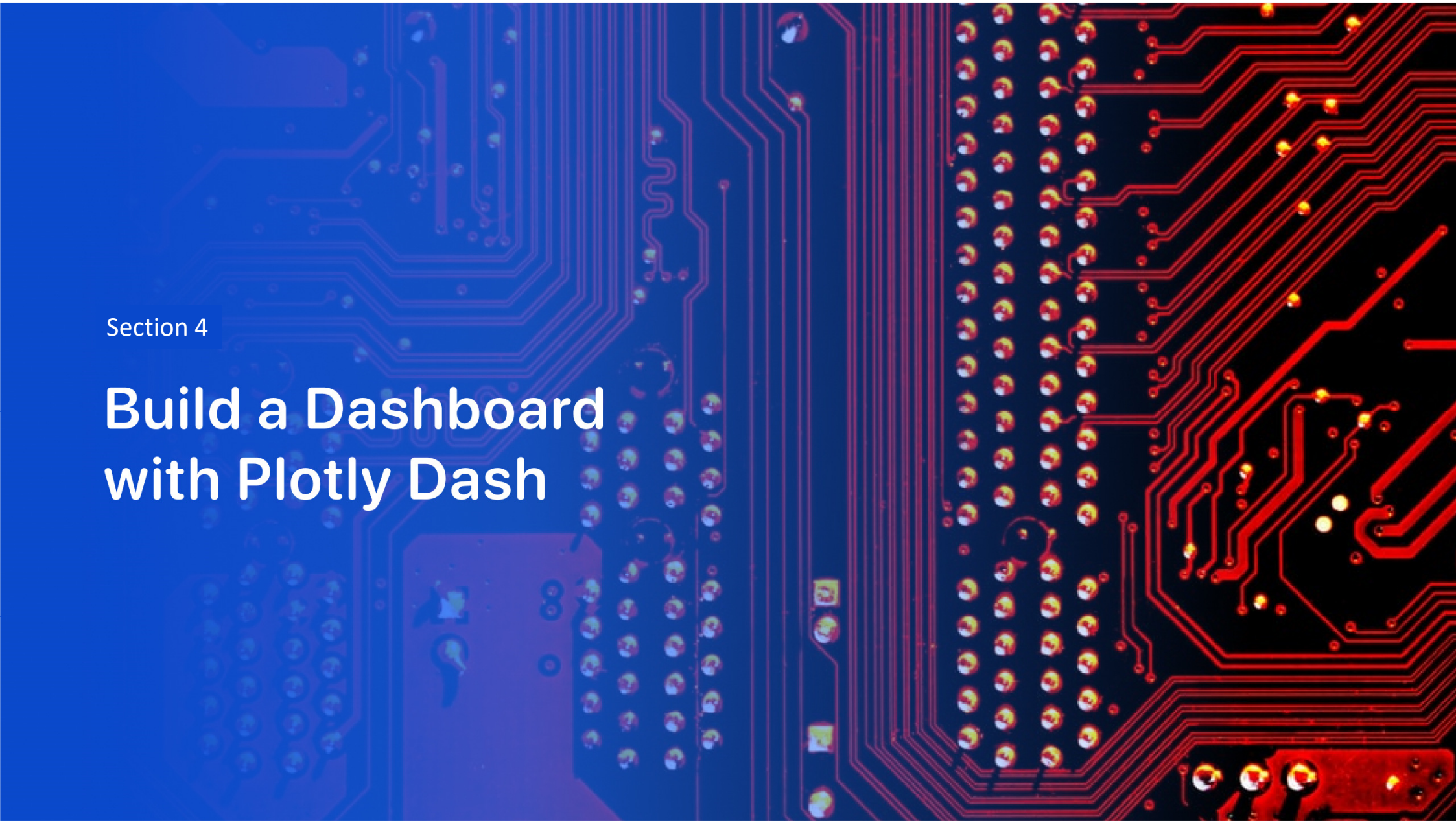


- Green markers indicate successful and red ones indicate failure.

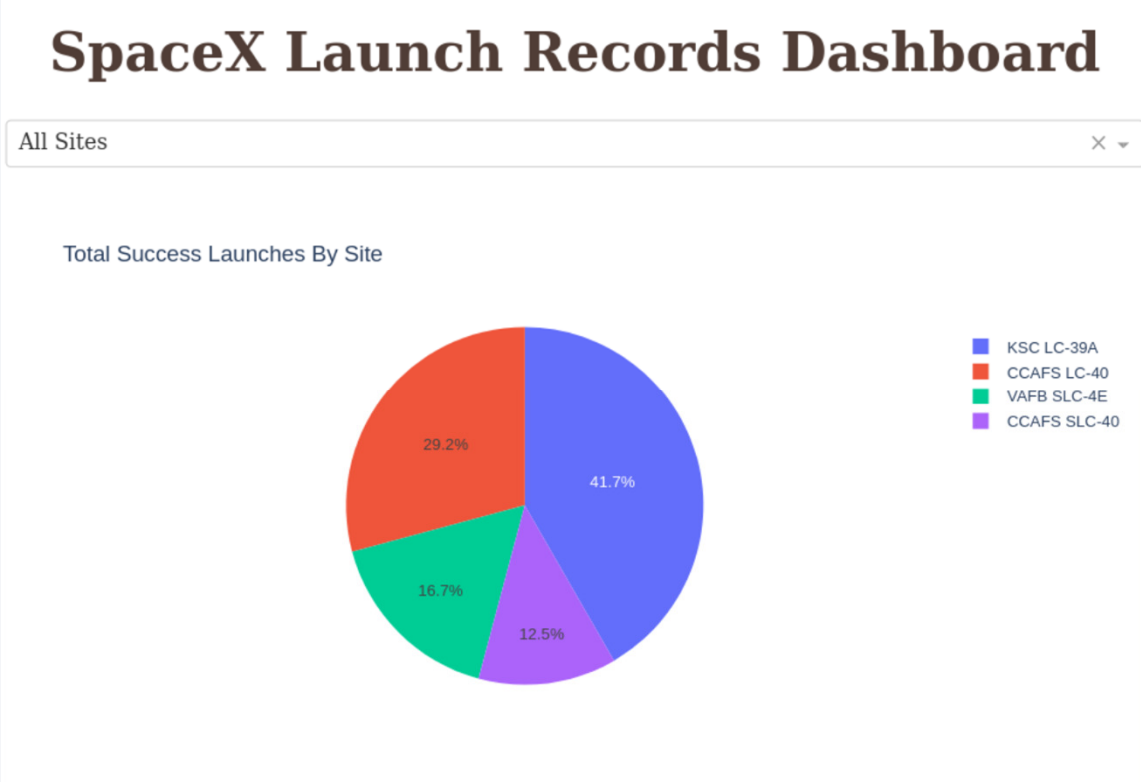
Logistics and safety



- CCAFS LC-40 has great logistics aspects being near to the coast, the railway line and the highway;
- Unfortunately is a relatively near to inhabited areas

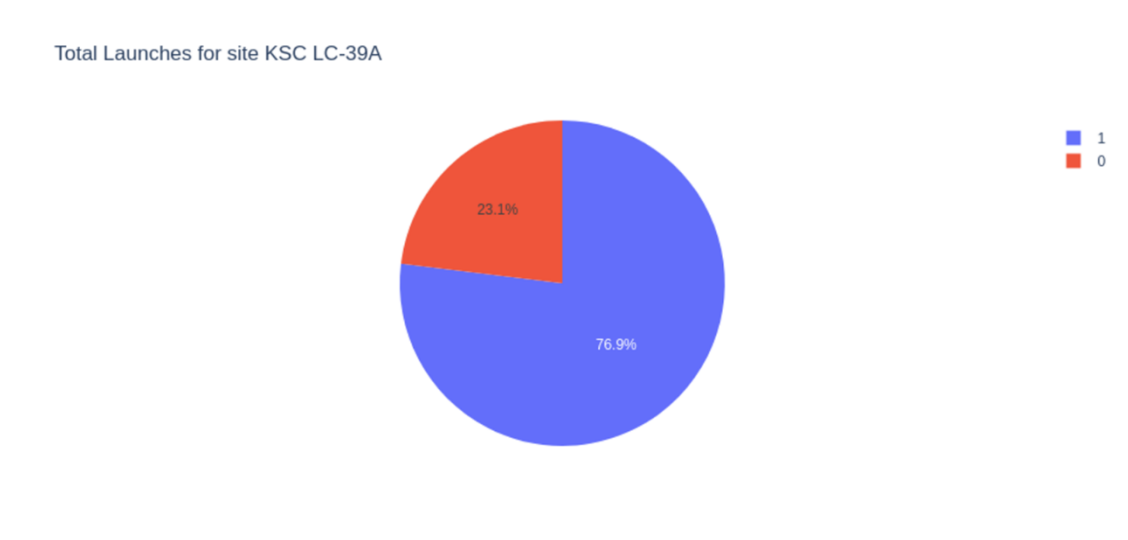

Section 4
Build a Dashboard with Plotly Dash

Successful launches by site



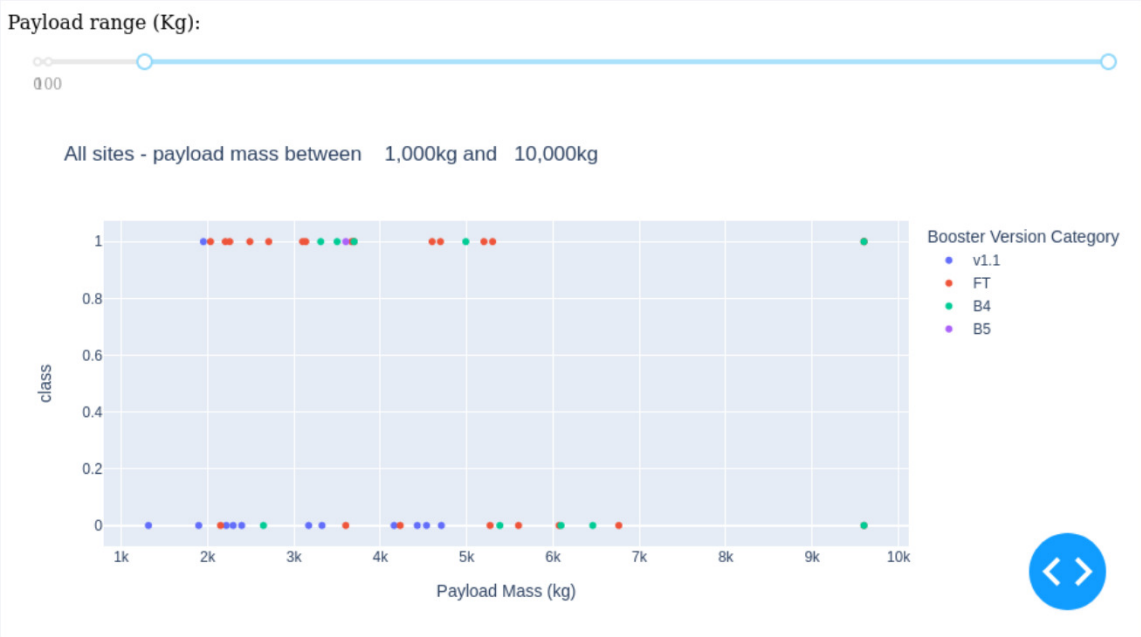
- Launch site seems to be an important factor of mission success

Launch success rate for KSC LC-39A



- 76.9% success rate!

Paylod vs. launch outcome



- Payloads under 6.000 kg and FT boosters are the most successful combination

Section 5

Predictive Analysis (Classification)

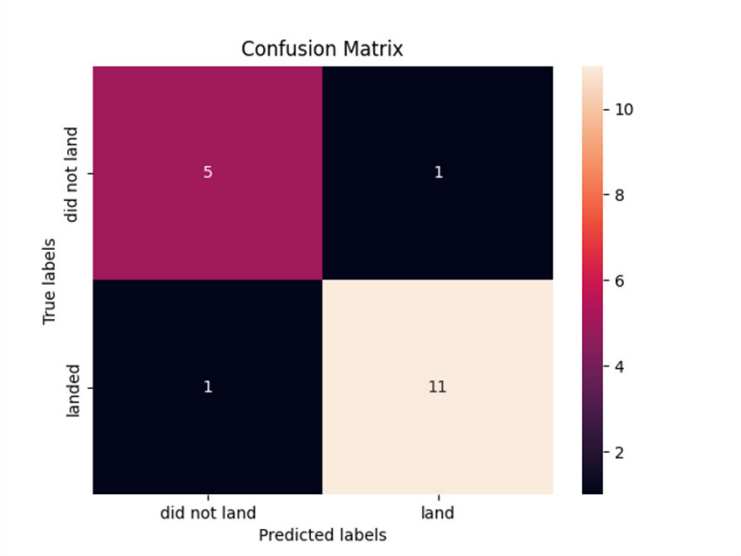
Classification Accuracy

- Four classification models were tested, accuracy listed here beside

- The highest classification accuracy was obtained with decision trees classifier

Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.90536	0.88889
KNN	0.84821	0.83333

Confusion Matrix



- True positive and true negative are very big compared to the false one, this proves the very good accuracy of Decisione Trres Classifier

Conclusions

- Different data sources analized with different methods
- KSC LC-39A is the best launch site
- Launches above 7.000 kg are less risky
- Success landing rate increases over time, highlighting the iterative approach typical of SpaceX; this is proof of processes and rockets evolution
- Decision tree classifier can be used to predict landing outcome and increase profits

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!