IBM Developer
SKILLS NETWORK

WINNING SPACE RACE WITH DATA SCIENCE

JIAWEI LI

13.02.2023

OUTLINE

1. Excutive Summary
2. Introduction
3. Methodology
4. Results
5. Conclusion
6. Appendix

EXECUTIVE SUMMARY

Summary of methodologies

- Data Collection
- Data Wrangling
- EDA with Data Visualization
- EDA with SQL
- Interactive Visual Analytics with Folium lab
- Create Dashboard with Plotly
- Prediction with Machine Learning

Summary of all results

- Results of EDA
- Demo for interactive Visual Analystivs
- Results of prediction

INTRODUCTION

Background

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. Therefor if we can determine if the first stage will land, we can determine the cost of a launch. With public information and machine learning models we can predict, if SpaceX will reuse the first stage.

Problems

- The relationships between variables and how they can have impact on outcome.
- What are the best conditions for the successful landing.
- Which machine learning models are better for this case.

METHODOLOGY

METHODOLOGY

Executive Summary

- Data collection methodology:
  - SpaceX REST API
  - Web Scrapping from Wikipedia
- Perform data wrangling
  - Data clean
  - Use One-Hot-Encoding for numeric variables
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - build, tune, evaluate classification models

METHODOLOGY

Data Collection

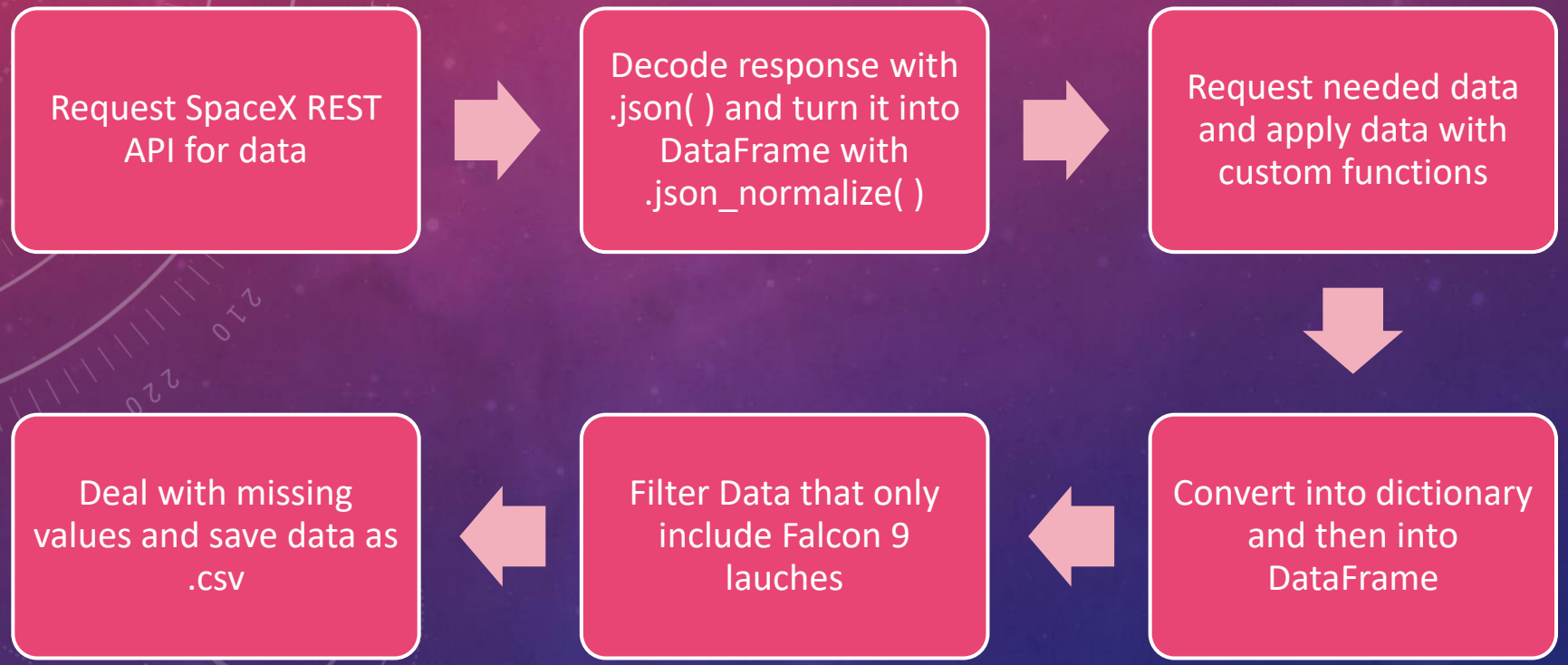
Data were collected from two sources:

- SpaceX REST API:
- https://api.spacexdata.com/v4/rockets/

- Wikipedia for SpaceX:
  https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches/

METHODOLOGY

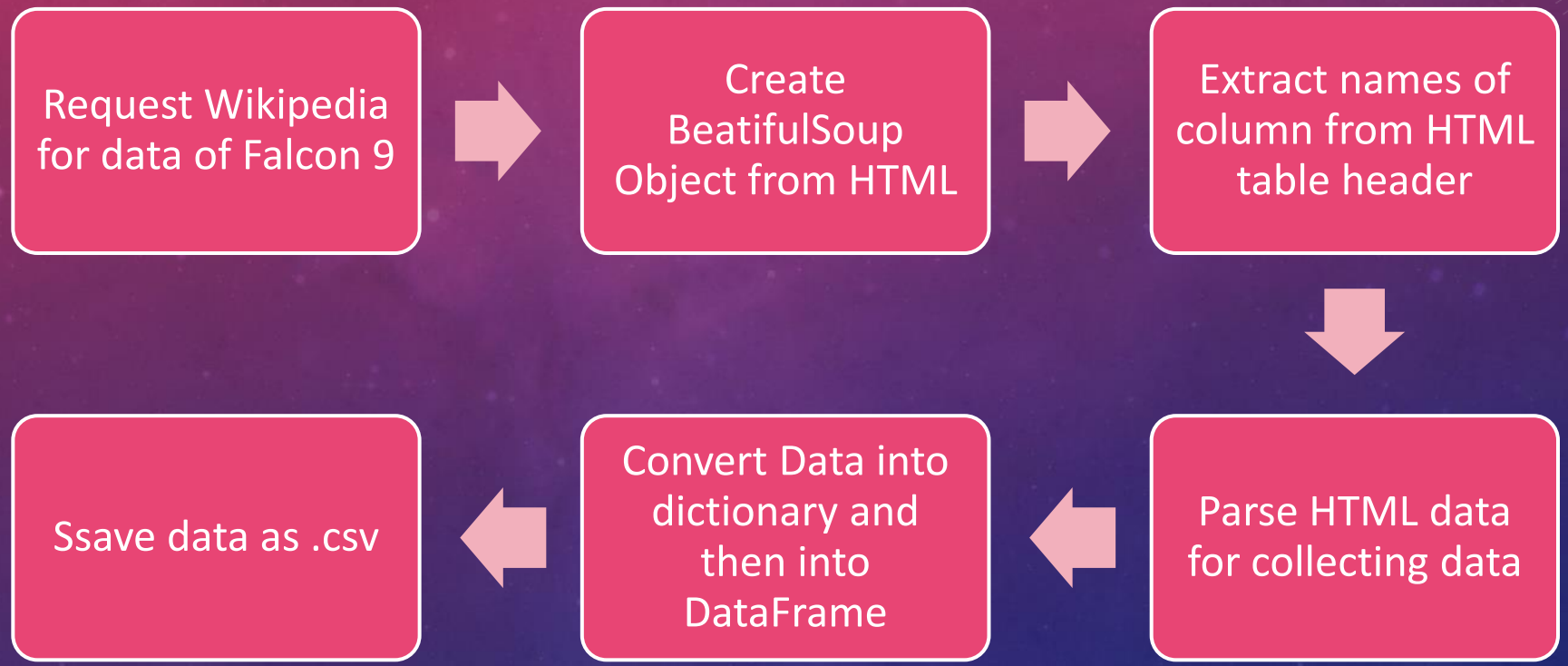
SpaceX REST API



Github: Data Collection API

METHODOLOGY

Wikipedia



Github: Data Collection with Web Scraping

METHODOLOGY

Data Wrangling

- EDA: Performe EDA to determine the the labels for training.
- Summarizations: Calculate the number of launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type.
- Create labels for data set: create labels from Outcome column.

Github: Data Wrangling

METHODOLOGY

EDA with Data Visualization

Variables	Chart Type
Flight Number & Payload Mass	Scatter
Flight Number & Launch Site	Scatter
Payload Mass & Launch Site	Scatter
Oribit & Class	Bar
Flight Number & Oribit	Scatter
Payload Mass & Oribit	Scatter
Year & Success Rate	Line

- Scatter: shows relationship between variables.
- Bar: compares discrete categories.
- Line: shows trends over time.

Github: EDA with Data Visualization

METHODOLOGY

EDA with SQL

- Names of the unique launch sites in missions
- 5 records where name of launch sites begins with ‘CCA’
- Total payload mass carried by boosters launched by NASA
- Average payload mass carried by booster version F9 v1.1
- Date of the first successful landing outcome
- Names of the successful boosters in drone ship and payload mass is greater than 4000 but less than 6000
- Total number of successful and failure mission outcomes
- Names of the booster versions that carried the maximum payload mass
- Failed landing outcomes in drone ship, which have booster versions and launch site names for the months in 2015
- Ranking the total number of landing outcomes between 2010-06-04 and 2017-03-20

Github: EDA with SQL

METHODOLOGY

Interactive Visual Analytics with Folium

Used tools:

- Markers indicates points (latitude and longitude coordinate)
- Circles indicates highlighted places around specific coordinates
- Marker clusters: indicates groups of places in each coordinate
- Lines: indicates distances between two coordinates

Github: Interactive Visual Analytics with Folium

METHODOLOGY

Interactive Dashboard with Ploty Dash

Success launches for sites:

- Dropdown list for sites selection
- Pie Chart for success und unsuccess Launches

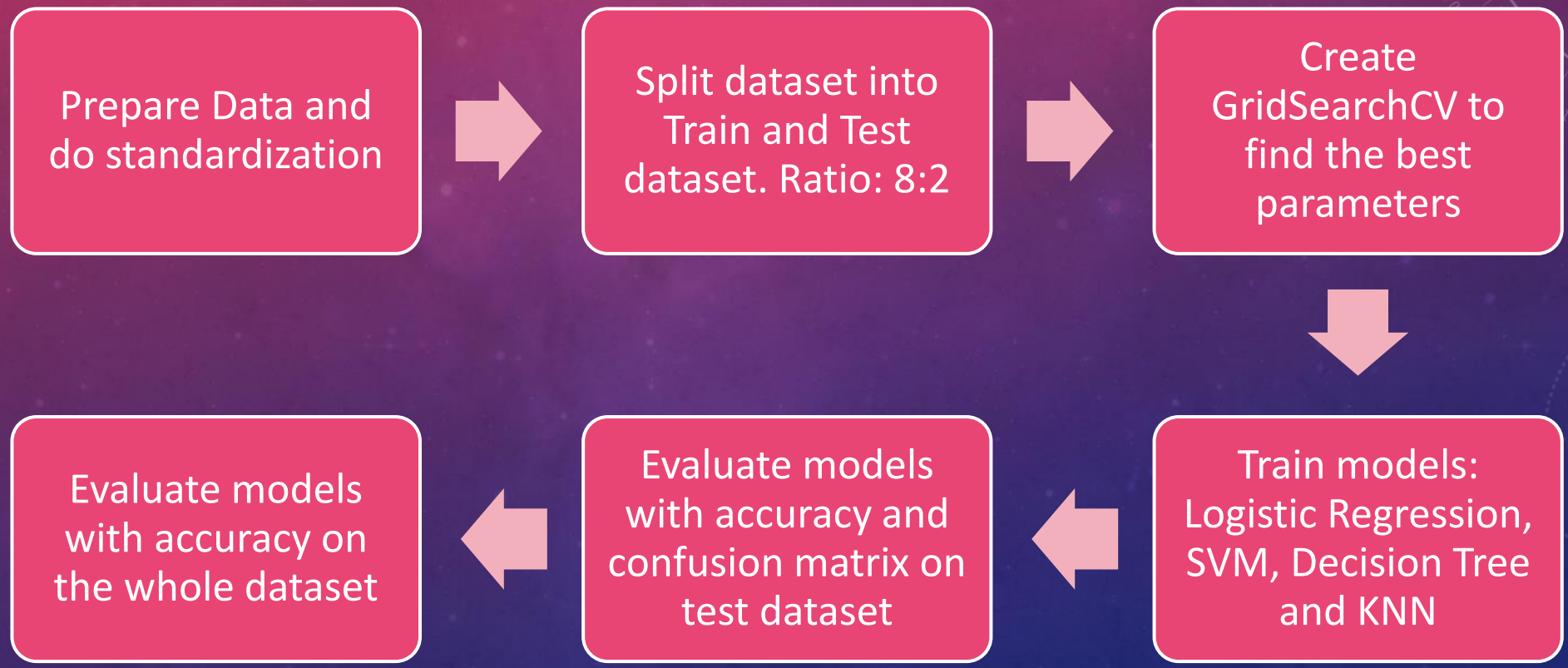
Correlation between payload mass and success:

- Payload mass range is 0 to 10000 kg
- Scatter Chart for correlation between payload mass and success

Github: Interactive Dashboard with Ploty Dash

METHODOLOGY

Machine Learning Prediction

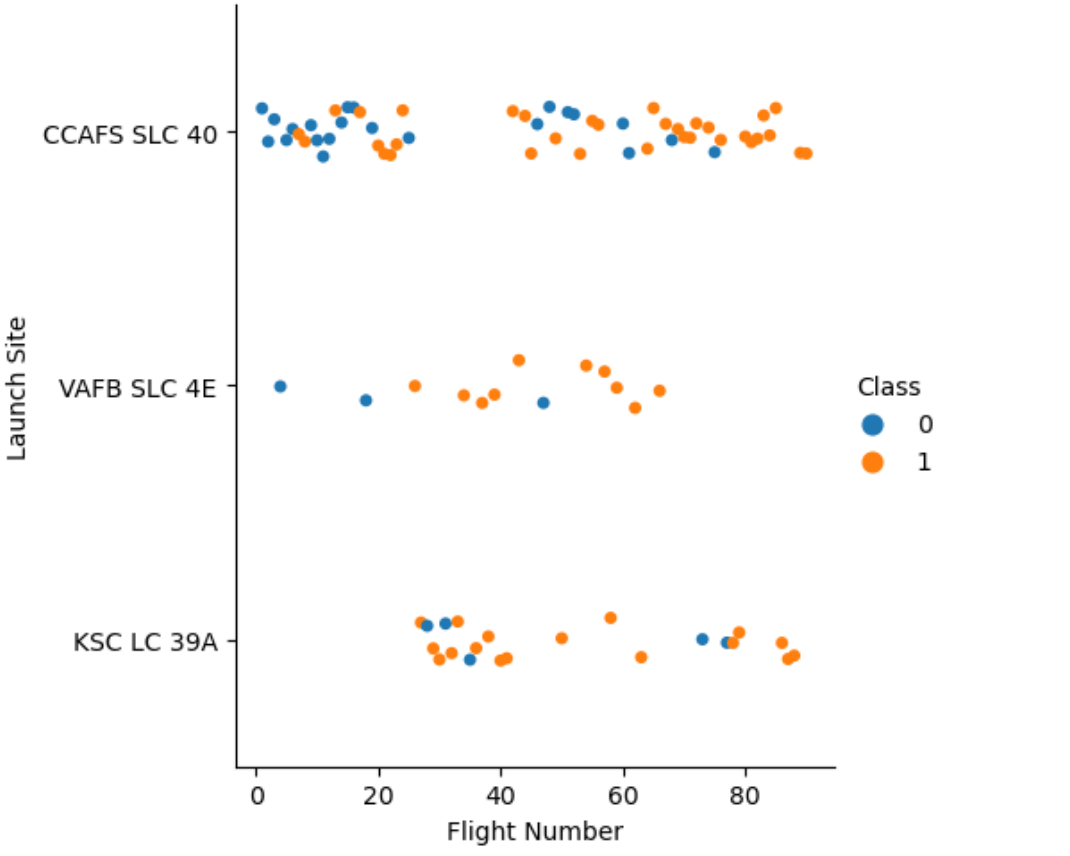


Github: Machine Learning Prediction

RESULTS

REULSTS

EDA with Data Visualization

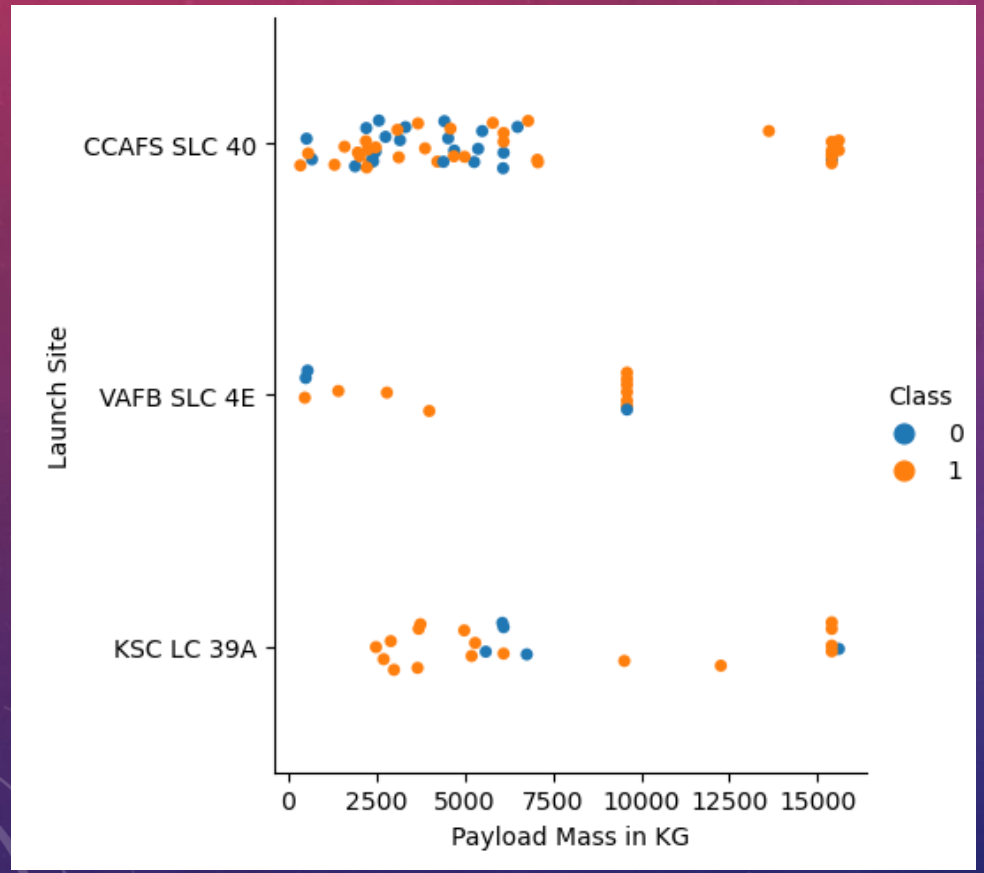


Flight Number and Launch Site:

- CCAFS SLC 40 has the most launches.
- KSC LC 39A has the highest success rate: 77.3%

REULSTS

EDA with Data Visualization

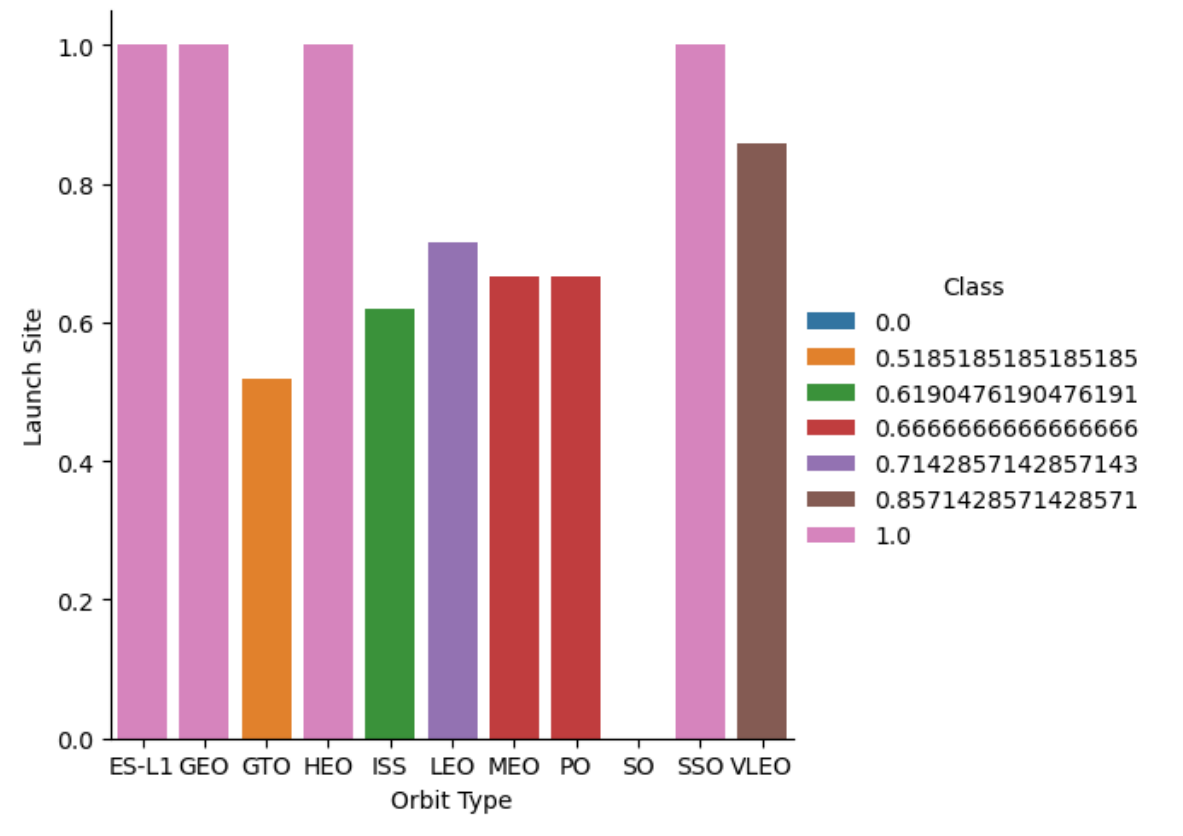


Payload mass and Launch Site:

- Higher payload mass, higher success rate
- CCAFS SLC 40 has 100% success rate when payload mass over 13000 kg
- KSC LC 39A has 100% success rate when payload mass under 5000 kg

REULSTS

EDA with Data Visualization

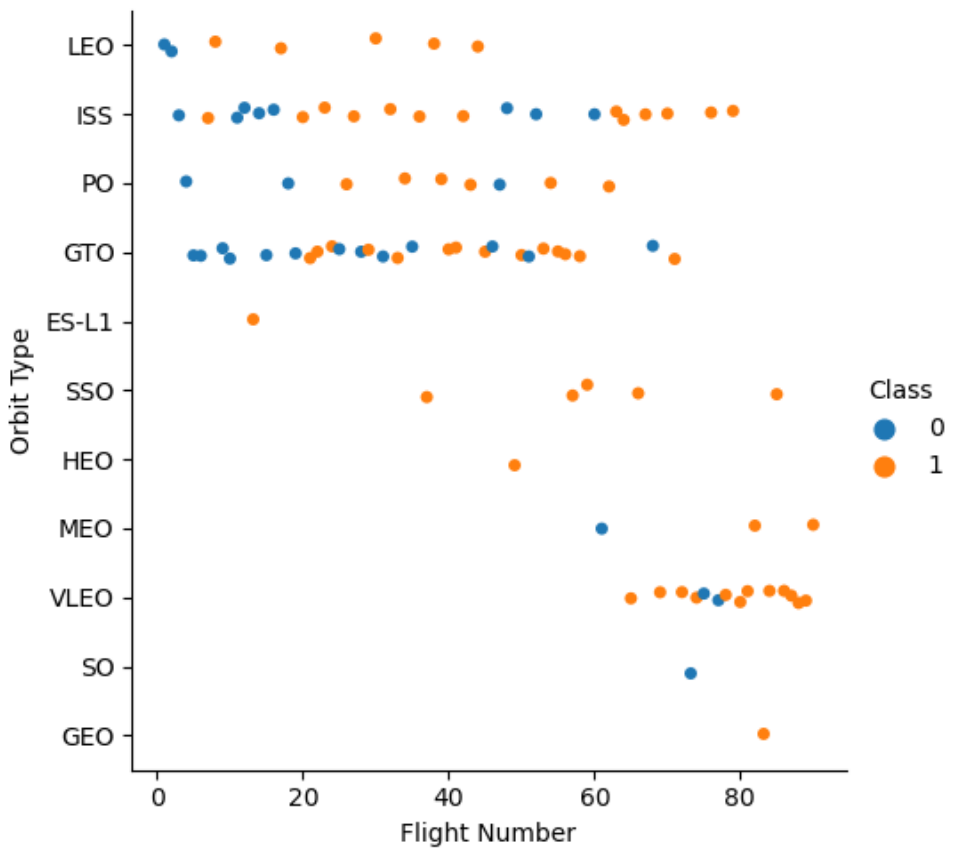


Orbit type and Launch Site:

- 4 Orbits has 100% success rate: ES-L1, GEO, HEO, SSO
- Orbit SO has 0% success rate

REULSTS

EDA with Data Visualization

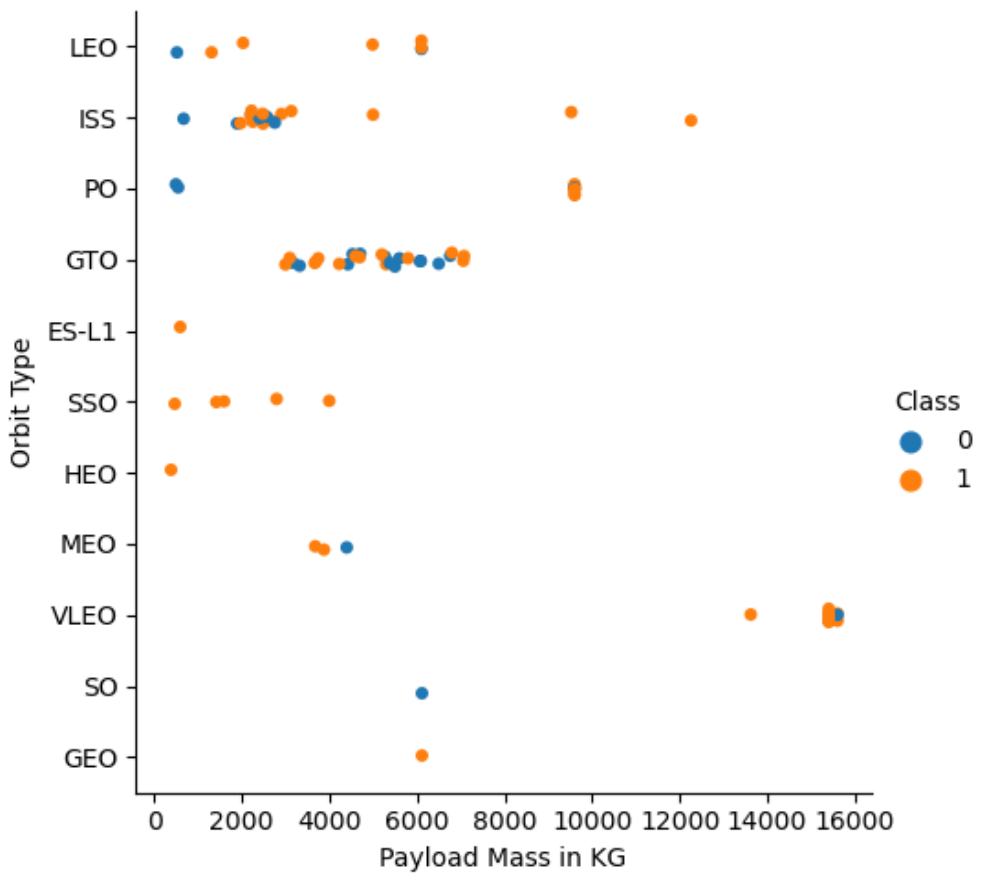


Flight number and Orbit type:

- Assume that flight number has no impact on Orbit ISS and GTO

REULSTS

EDA with Data Visualization

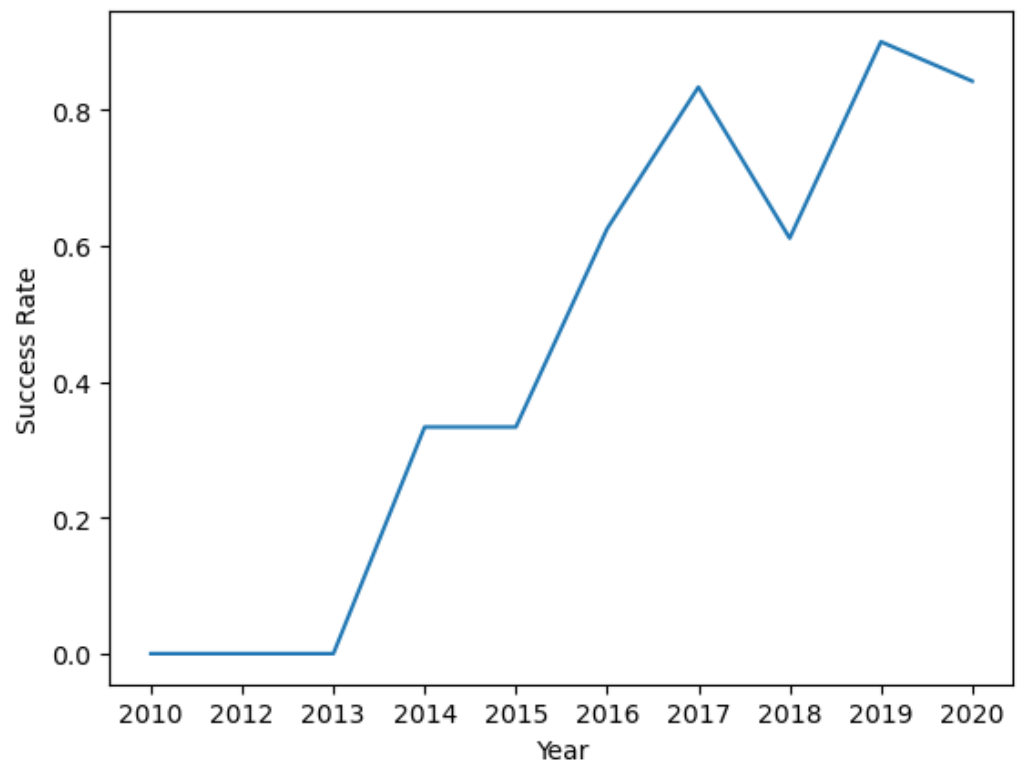


Payload mass and Orbit type:

- Assume that payload mass has no impact on Orbit GTO

REULSTS

EDA with Data Visualization



Year and Success rate:

- Success rate kept increasing after 2013

REULSTS

EDA with SQL

Names of the unique launch sites in missions:

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

5 records where name of launch sites begins with ‘CCA’ :

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

REULSTS

EDA with SQL

Total payload mass carried by boosters launched by NASA:

total_payload_mass
45596

Average payload mass carried by booster version F9 v1.1:

average_payload_mass
2534.6666666666665

Date of the first successful landing outcome:

first_successful_landing
01-05-2017

REULSTS

EDA with SQL

Names of the successful boosters in drone ship and payload mass is greater than 4000 but less than 6000:

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

Total number of successful and failure mission outcomes:

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

REULSTS

EDA with SQL

Names of the booster versions that 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

REULSTS

EDA with SQL

Failed landing outcomes in drone ship, which have booster versions and launch site names for the months in 2015:

month	Date	Booster_Version	Launch_Site	Landing _Outcome
01	10-01-2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	14-04-2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

REULSTS

EDA with SQL

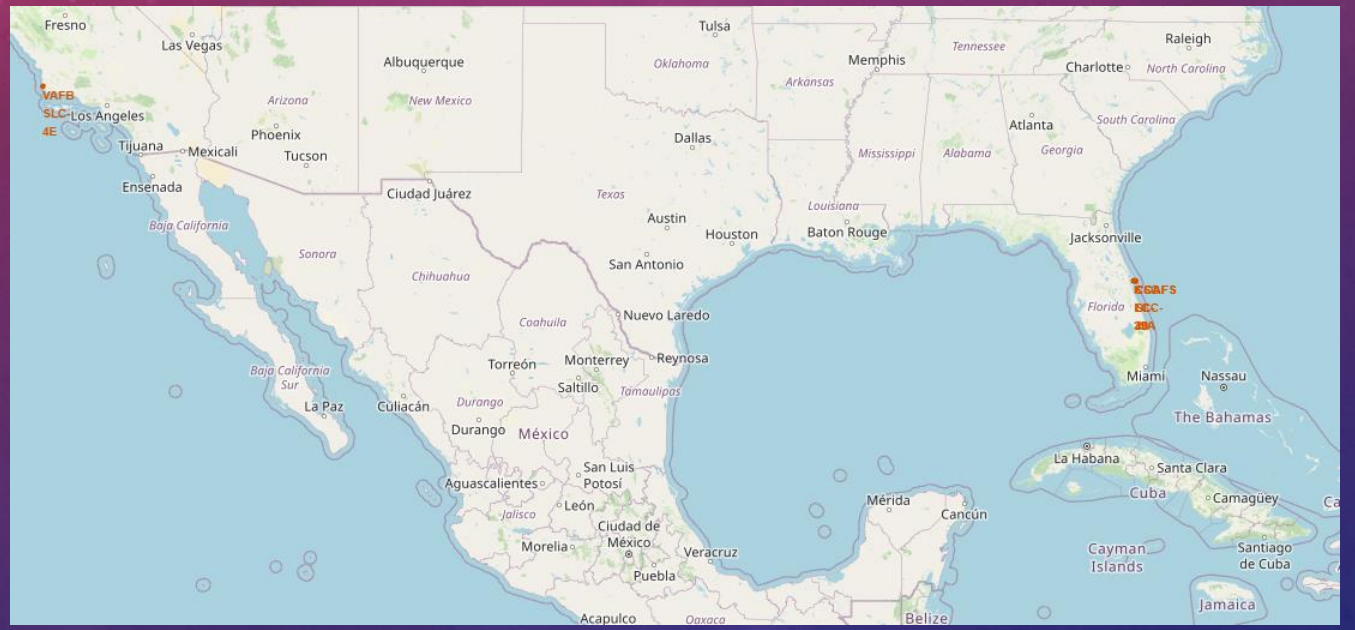
Ranking the total number of landing outcomes between 2010-06-04 and 2017-03-20

Landing _Outcome	count_outcomes
Success	24
No attempt	14
Success (ground pad)	8
Success (drone ship)	8
Failure (drone ship)	5
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

REULSTS

Interactive Visual Analytics with Folium

Location of all launch sites on a global map:

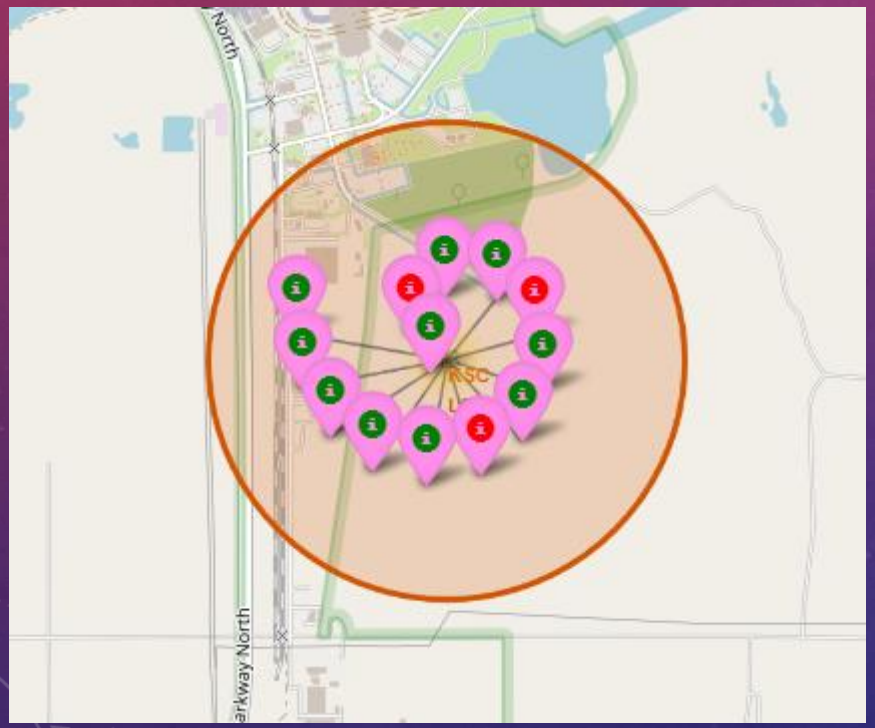


- All launch sites are very close proximity to the coast.
- Most of Launch sites are close to the Equator line.

REULSTS

Interactive Visual Analytics with Folium

Launch records on the map:

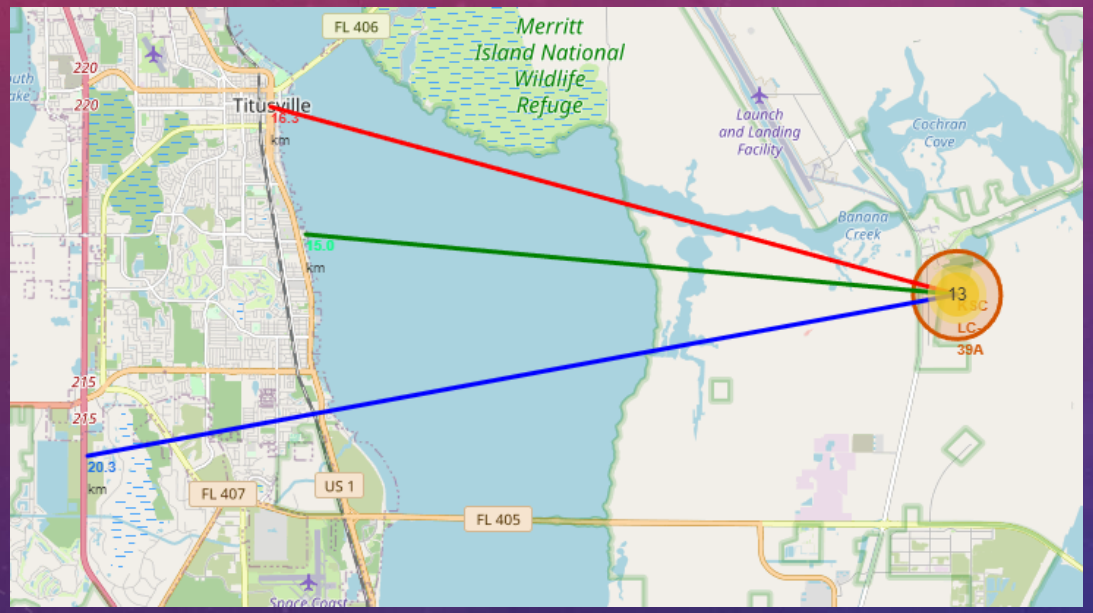


- Green marker: successful launch
  red marker: unsuccessful launch
- KSC LC-39A has a very high Success Rate. (This left picture shows KSC LC-39A)

REULSTS

Interactive Visual Analytics with Folium

Distance from KSC LC-39A:

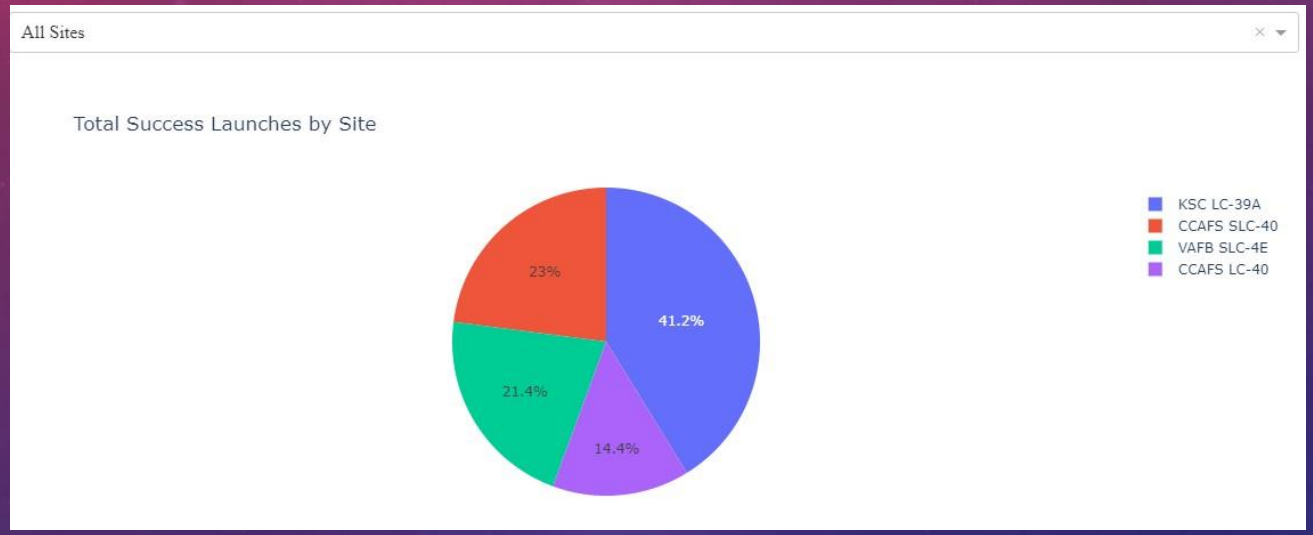


- Distance to city Titusville: 16.32 km
- Distance to coastline: 15.8 km
- Distance to highway: 20.3 km

REULSTS

Interactive Dashboard with Ploty Dash

Launch success for all sites:

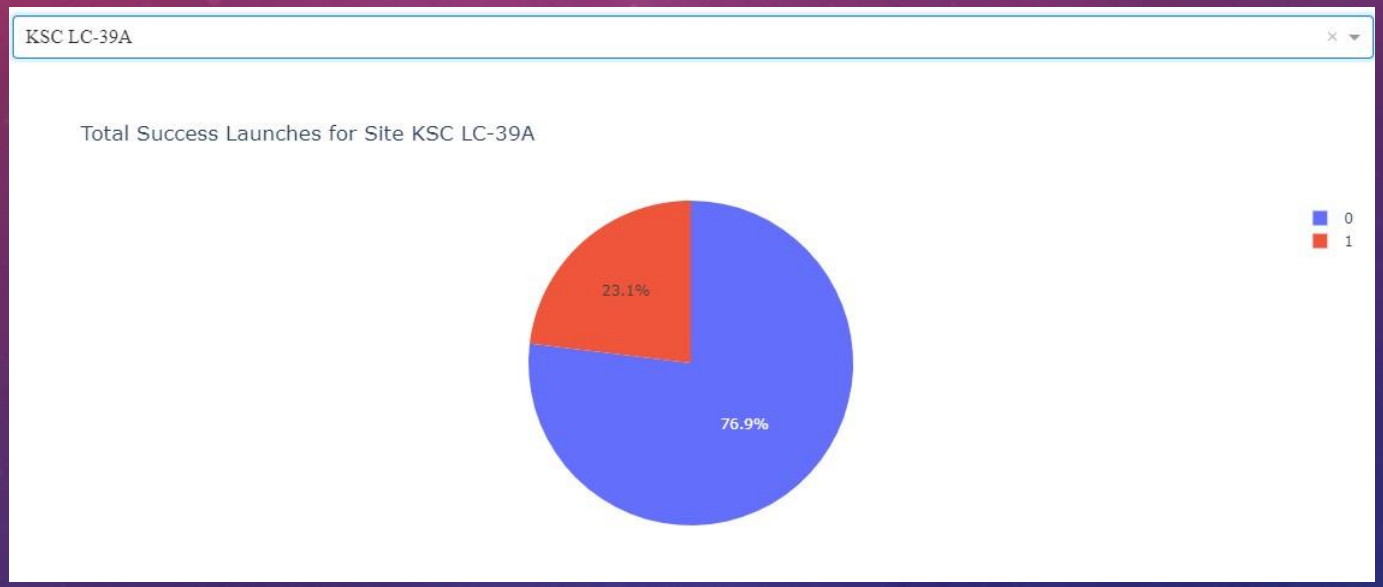


- KSC LC-39A has the most success
- CCAFS LC-40 has the least success

REULSTS

Interactive Dashboard with Ploty Dash

highest launch success ratio:

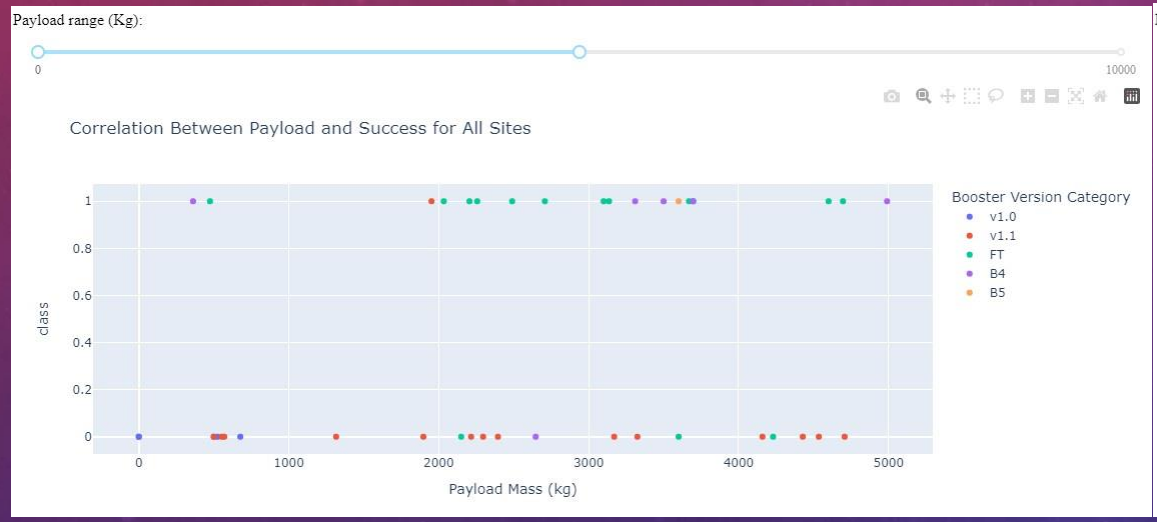


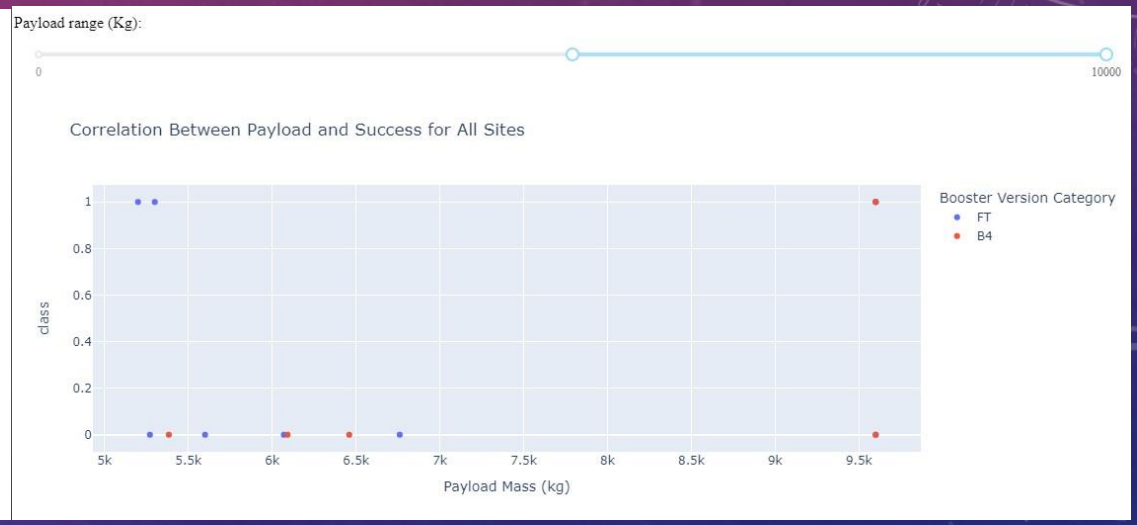
- KSC LC-39A has the highest success ratio with 76.9%

REULSTS

Interactive Dashboard with Ploty Dash

Payload mass and launch outcome for all sites:





- payloads between 2000 and 6000 kg have the highest success rate.

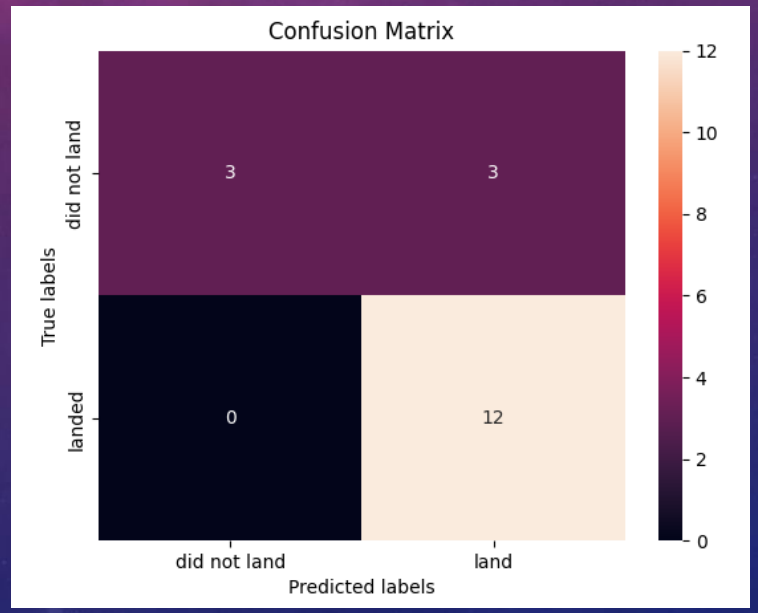
REULSTS

Machine Learning Prediction

Training dataset: Test dataset = 8:2

Test dataset: 18 samples

Confusion matrix of models are all same on test dataset:



REULSTS

Machine Learning Prediction

Accuracy of all models on test dataset are all 93.3%

To find best models, accuracy of all models on the whole dataset are calculated:

Model	Logistic Regression	SVM	Decision Tree	KNN
Accuracy	86.7%	87.8%	81.1%	85.6

So, the best model is SVM.

CONCLUSION

CONCLUSION

- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.
- KSC LC-39A has the highest success rate.
- Success rate of launches kept increasing over the years.
- All sites are very close to the coast.
- Low payload mass performances better.
- SVM is the best model for this dataset.

APPENDIX

APPENDIX

Courseara

Instructors

IBM

THANK YOU!