IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

a Data Science Capstone Project

by
Boris Nedyalkov
09/05/2023

Outline

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

Executive Summary

- **Summary of methodologies**
  - Data from all the launches of Space X in the last 16 years was scraped with Space X's API
  - The data was organized in pandas dataframe, cleaned, classified and prepared for machine learning processing
  - SQL queries and pandas commands were used for the data processing and to acquire insights from the data
  - Seaborn was used to deduce and extract plots of valuable information and acquire insights from the data
  - Folum was used to describe visually the launch sites and outcomes on a map
  - Plotly and Dash were used to create interactive dashboards to illustrate the data
  - Machine learning models were created to predict the success/failure of each Space X mission

- **Summary of all results**
  - Linear increasing correlation is established b/w incidencies of successful landings vs time
  - The most used orbits and successful launch sites are pointed out
  - A model with an accuracy rate of ~77% was established, that predicts the probability of successful landing of Space X rockets

Introduction

Problems

- How can Space Y learn from Space X's trials
- How can one predict the outcomes of the launches with a machine learning model

Background

- The Commercial Space business is an upcoming niche
- Space X has the best pricing ($62 million vs. $165 million compared to NASA)
- Mostly due to the ability to recover the first stage of the rocket
- Space Y would like to to compete with Space X

Falcon 9 rocket

Section 1

Methodology

Methodology

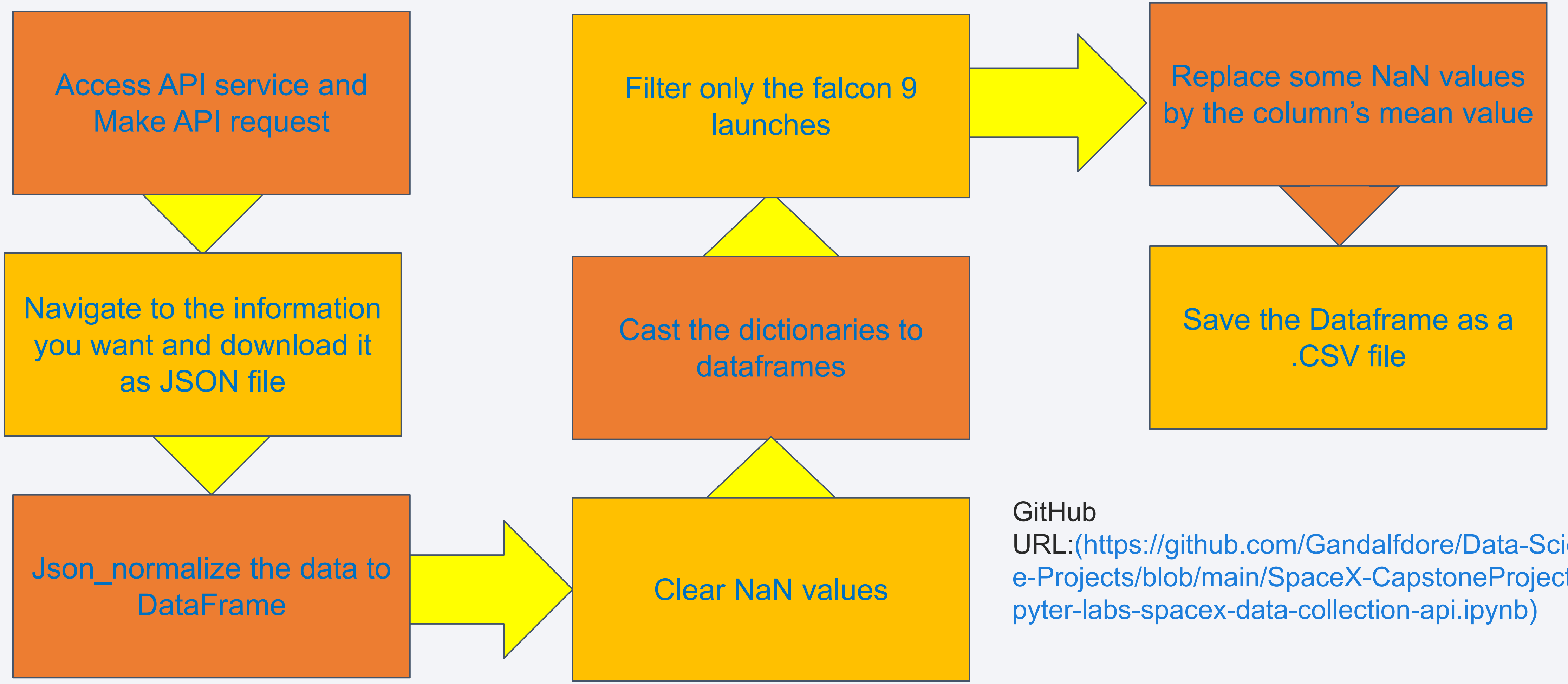
Executive Summary

- Data collection methodology:
  - Data taken from SpaceX's public API
- Performed data wrangling
  - Trough SQL queries and Pandas sorting feature engineering
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - Classification models built (KNN, Regression Trees, SVM and Logistic Regression)
  - The models' hyperparameters were tuned by Grid Search and the best model was selected (the SVM)

Data Collection

- **Collection**
  - The data was collected using requests to the REST Space X API
  - The API: "https://api.spacexdata.com/v4/rockets/"
  - The data was taken in the form of .JSON files
- **The data collected was for 80 flights of Space X rockets**
- **The data included in it was information about:**
  - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins
  - Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Data Collection – SpaceX API



GitHub URL:(https://github.com/Gandalfdore/Data-Science-Projects/blob/main/SpaceX-CapstoneProject/jupyter-labs-spacex-data-collection-api.ipynb)

Data Wrangling

- The data was cleaned as such
  - Categorical data were clearly separated
  - The ‘Mission Outcome’ was classified into ‘Class’ column
  - ‘Class’ takes value ‘0’ for failure and ‘1’ for success
  - One hot encode the categorical variables

❑ True ASDS, True RTLS, & True Ocean – set to -> 1
❑ None None, False ASDS, None ASDS, False Ocean, False RTLS – set to -> 0

- GitHub: https://github.com/Gandalfdore/Data-Science-Projects/blob/main/SpaceX-CapstoneProject/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

- Exploratory Data Analysis performed on variables Flight Number, Payload Mass, Launch Site, Orbit, Class and Year
- Plots generated:
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Payload Mass vs. Launch Site
  - Orbit vs. Success Rate
  - Flight Number vs. Orbit
  - Payload vs Orbit
  - Success Yearly Trend

- The plots were scatter plots, bar plots and a line plot

[GitHub URL: https://github.com/Gandalfdore/Data-Science-Projects/blob/main/SpaceX-CapstoneProject/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb](https://github.com/Gandalfdore/Data-Science-Projects/blob/main/SpaceX-CapstoneProject/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb)

EDA with SQL

- SQL queries are used to get insights from the data

- SQLLite with Jupyter notebook was used for the purpose

- Queries were made about the launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes

  [GitHub URL: https://github.com/Gandalfdore/Data-Science-Projects/blob/main/SpaceX-CapstoneProject/jupyter-labs-eda-sql-coursera_sqllite.ipynb](https://github.com/Gandalfdore/Data-Science-Projects/blob/main/SpaceX-CapstoneProject/jupyter-labs-eda-sql-coursera_sqllite.ipynb)

Build an Interactive Map with Folium

- Folium map was created with:
  - Marks for the different Launch Sites
  - Successful and unsuccessful landings pin, each with different color
  - Proximity references to key locations: Railways, Highways, Coasts, and Cities

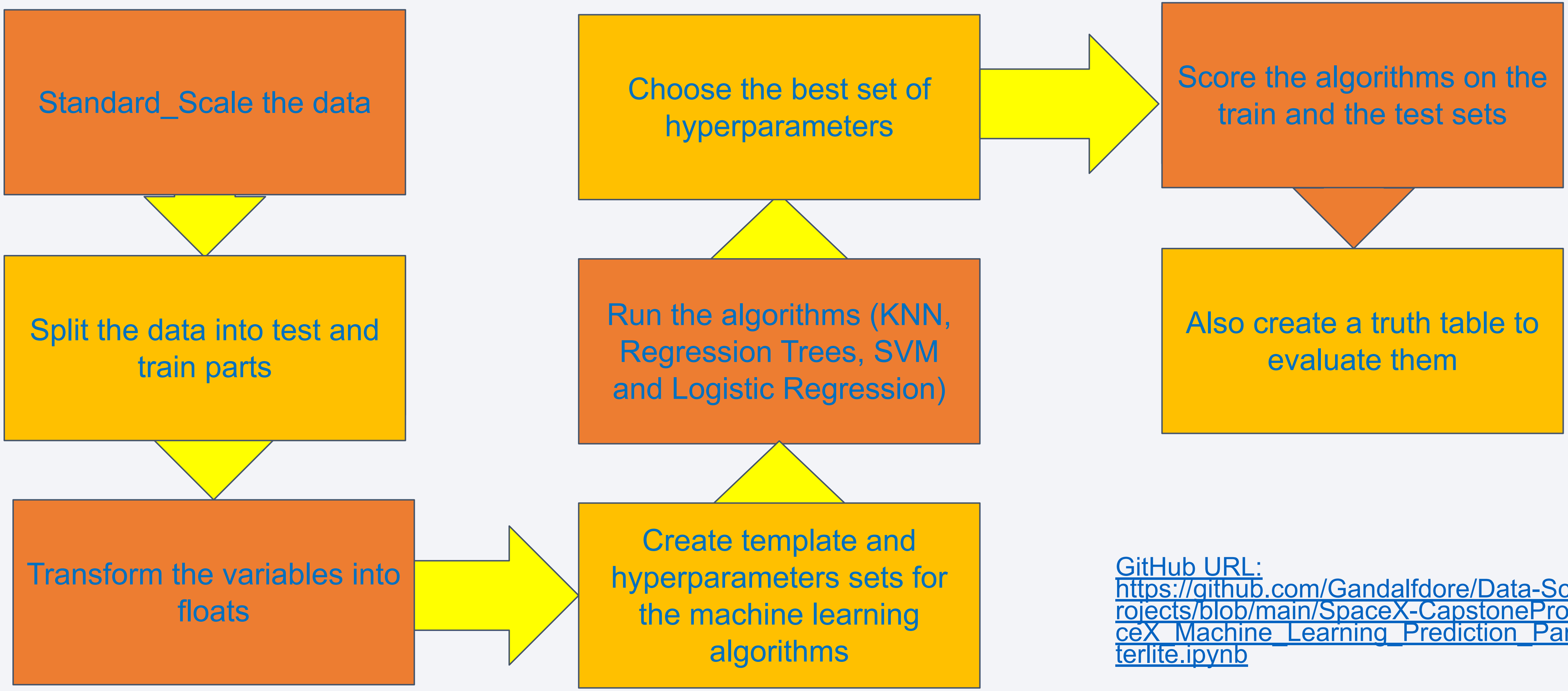
[GitHub URL: https://github.com/Gandalfdore/Data-Science-Projects/blob/main/SpaceX-CapstoneProject/folium_workshop.ipynb](https://github.com/Gandalfdore/Data-Science-Projects/blob/main/SpaceX-CapstoneProject/folium_workshop.ipynb)

Build a Dashboard with Plotly Dash

- The dashboard has:
  - Pie chart
  - Scatter plot.
- The pie chart shows:
  - Distribution of successful landings per launch site
  - Individual launch site success rates.
- Scatter (inputs: individual site and payload mass between 0 and 10000 kg)

  [GitHub URL: https://github.com/Gandalfdore/Data-Science-Projects/blob/main/SpaceX-CapstoneProject/spacex_dashboard_app.py](https://github.com/Gandalfdore/Data-Science-Projects/blob/main/SpaceX-CapstoneProject/spacex_dashboard_app.py)

Predictive Analysis (Classification)



GitHub URL: https://github.com/Gandalfdore/Data-Science-Projects/blob/main/SpaceX-CapstoneProject/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

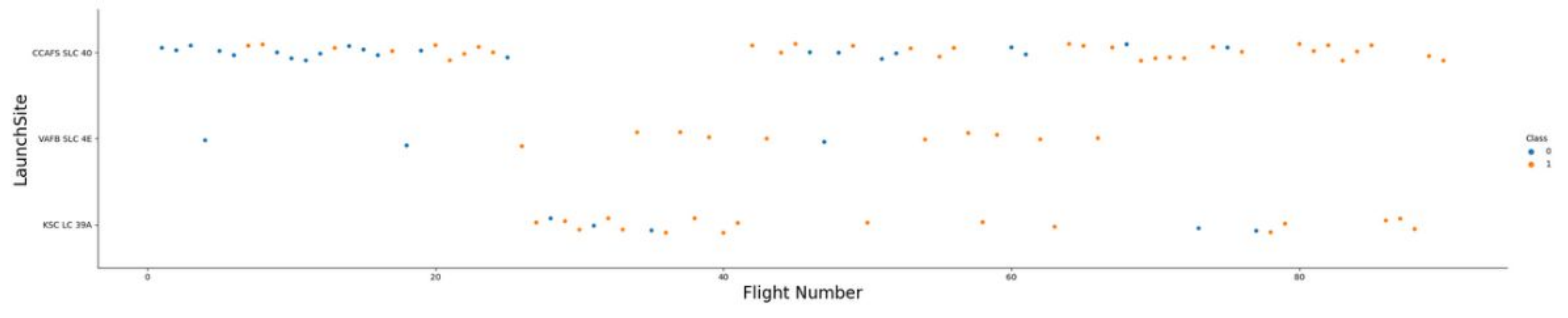
Results

- There is a clear trend between time passed and improvements in the success rates of rocket landing and launches
- Launches for certain orbits have greater success the for other orbits
- We create a model that can predict with ~77% accuracy the probability of successful landing of Space X rockets

Section 2

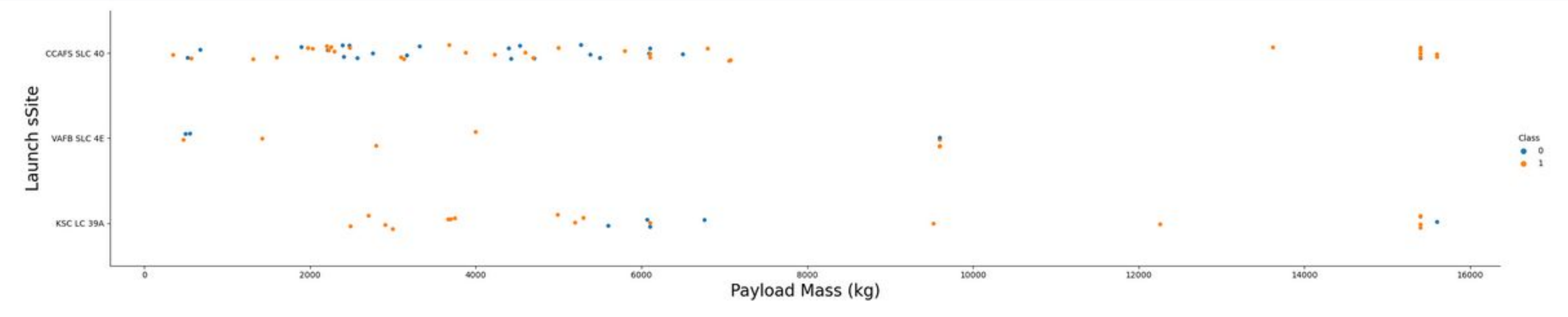
Insights drawn from EDA

Flight Number vs. Launch Site



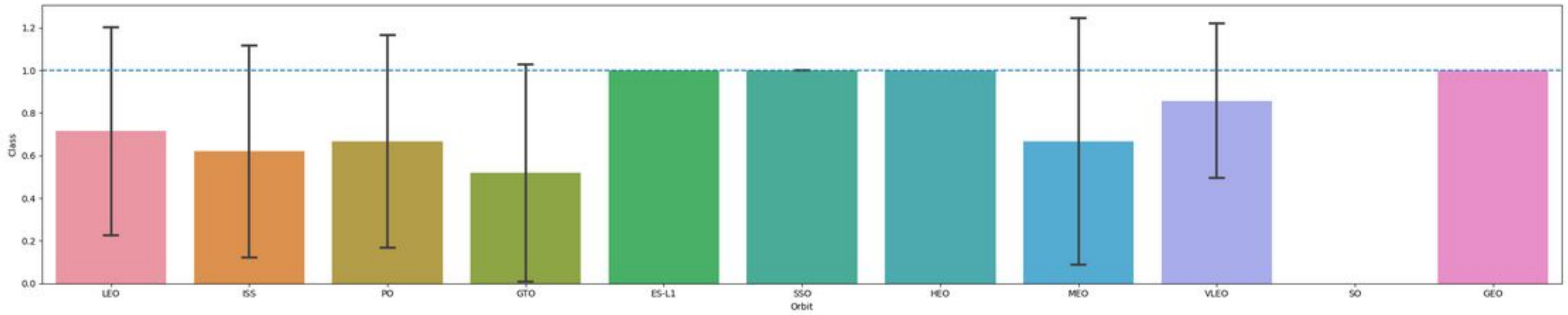
- ★ There is clearly more rockets launched from Cape Canaveral
- ★ The success rate of the flights seems to improve with time

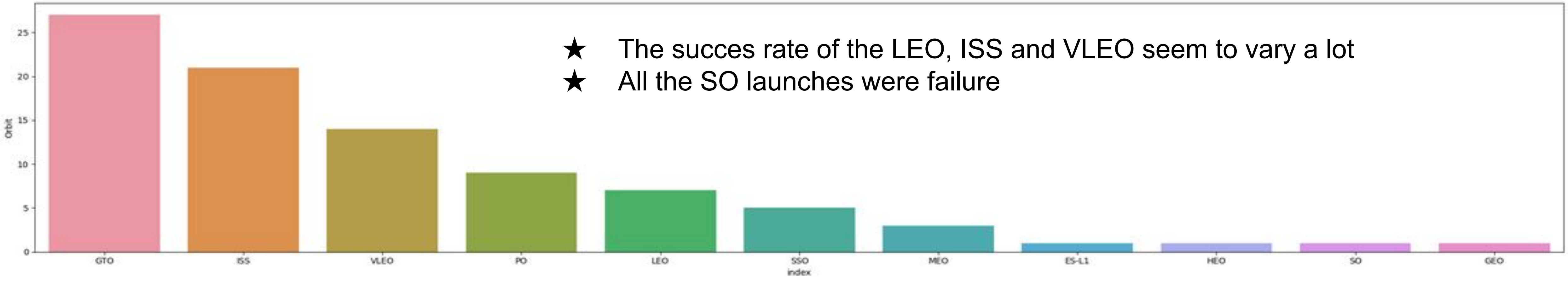
Payload vs. Launch Site



- ★ There are no rockets heavier than 10000 kg launched from VAFB SLC
- ★ Ships with bigger payloads have a higher tendency to be successfully landed and launched

Success Rate vs. Orbit Type

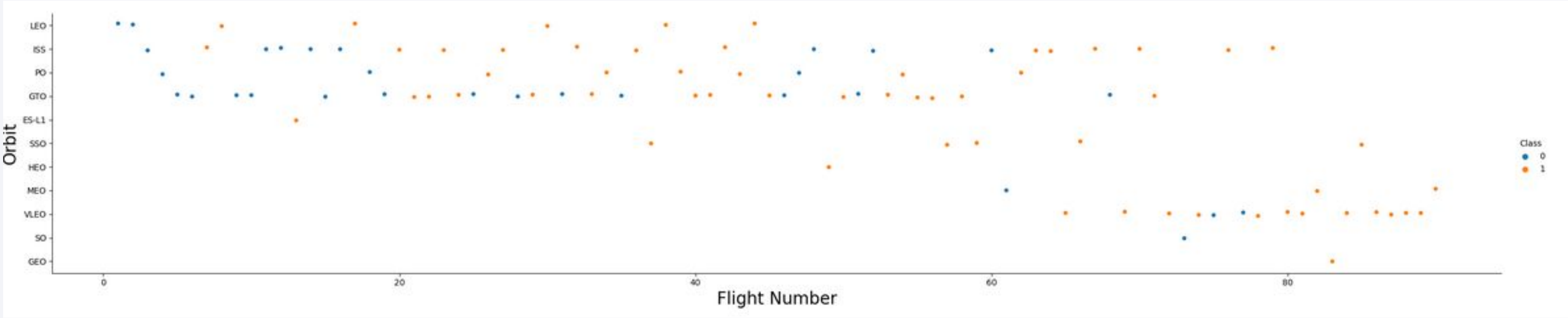




★ The succes rate of the LEO, ISS and VLEO seem to vary a lot
★ All the SO launches were failure

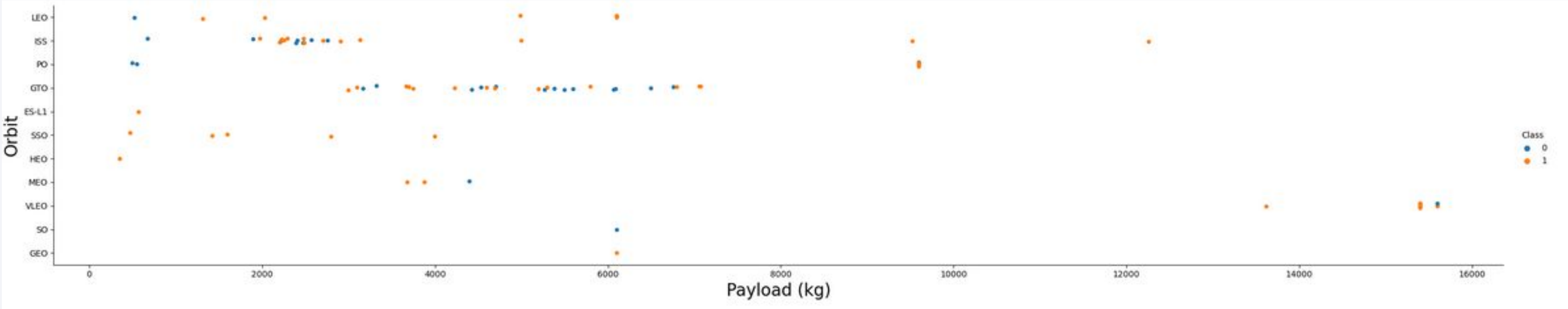
★ The success rate of the LEO, SSO, GEO and MEO orbits launches was 100%, but their samples are comparatively small

Flight Number vs. Orbit Type



- ★ For the LEO orbit the Success appears related to the number of flights
- ★ There seems to be no relationship between flight number when in GTO orbit
- ★ SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches
- ★ SpaceX appears to perform better in lower orbits or Sun-synchronous orbits

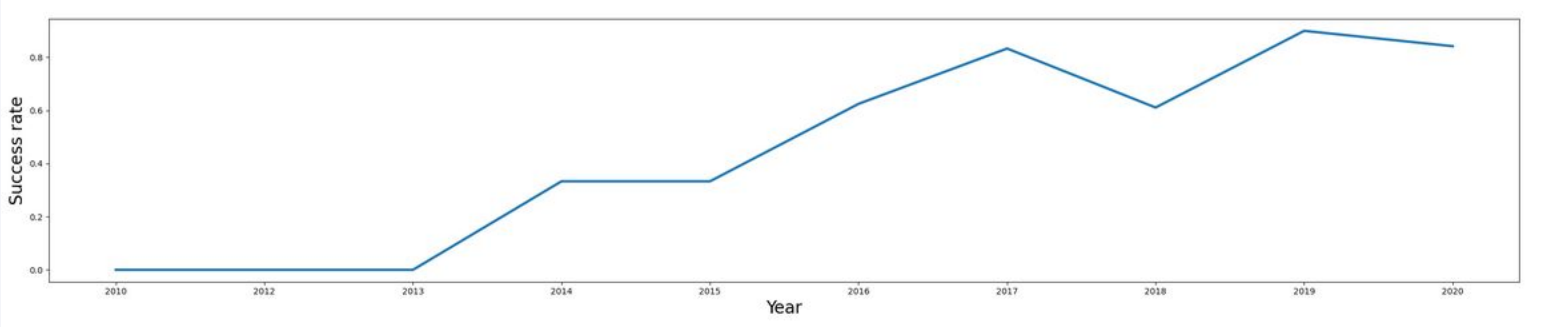
Payload vs. Orbit Type



★ With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

★ However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



- ★ It is clear that each year on average the success rates of the launches/landings is getting better
- ★ Success in recent years at around 80%

All Launch Site Names

- 3 unique launch sites
- CCAFS SLC-40 and CCAFS LC-40 are the same launch site

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

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outc
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (paracl
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 (paracl
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No atte
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No atte
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No atte

- First five entries in database with Launch Site name beginning with CCA.

Total Payload Mass

- The mass of all the payloads from the last 16 years

SUM(PAYLOAD_MASS__KG_)
99980

Average Payload Mass by F9 v1.1

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

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad of Space X rocket

MIN(Date)
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

- These are the names of boosters which have successfully landed on drone ship and have payload mass greater than 4000 kg but less than 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

- This is the total number of successful and failure mission outcomes
- The 2 Success entries are there by error actually they belong together ;)

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

Boosters Carried Maximum Payload

- These are the names of the booster which have carried the maximum payload mass

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

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

MONTH	landing__outcome	booster_version	payload_mass__kg_	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	2395	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	1898	CCAFS LC-40

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

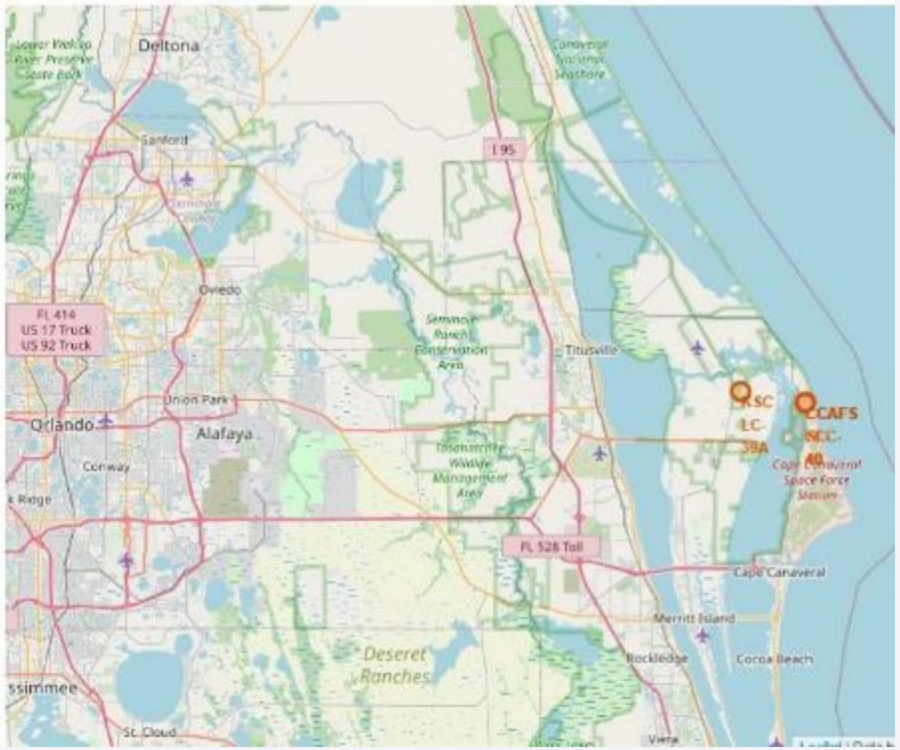
- Count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

landing__outcome	no_outcome
Success (drone ship)	5
Success (ground pad)	3

Section 3

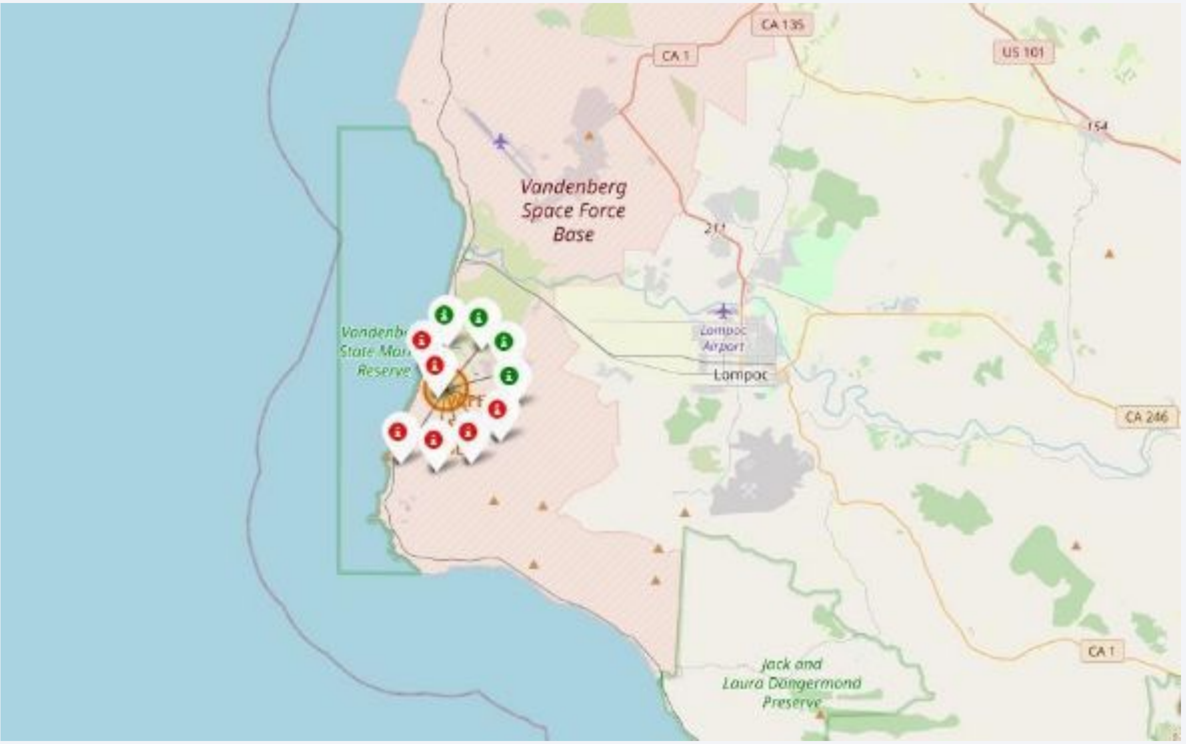
Launch Sites Proximities Analysis

<Folium Map Screenshot 1>



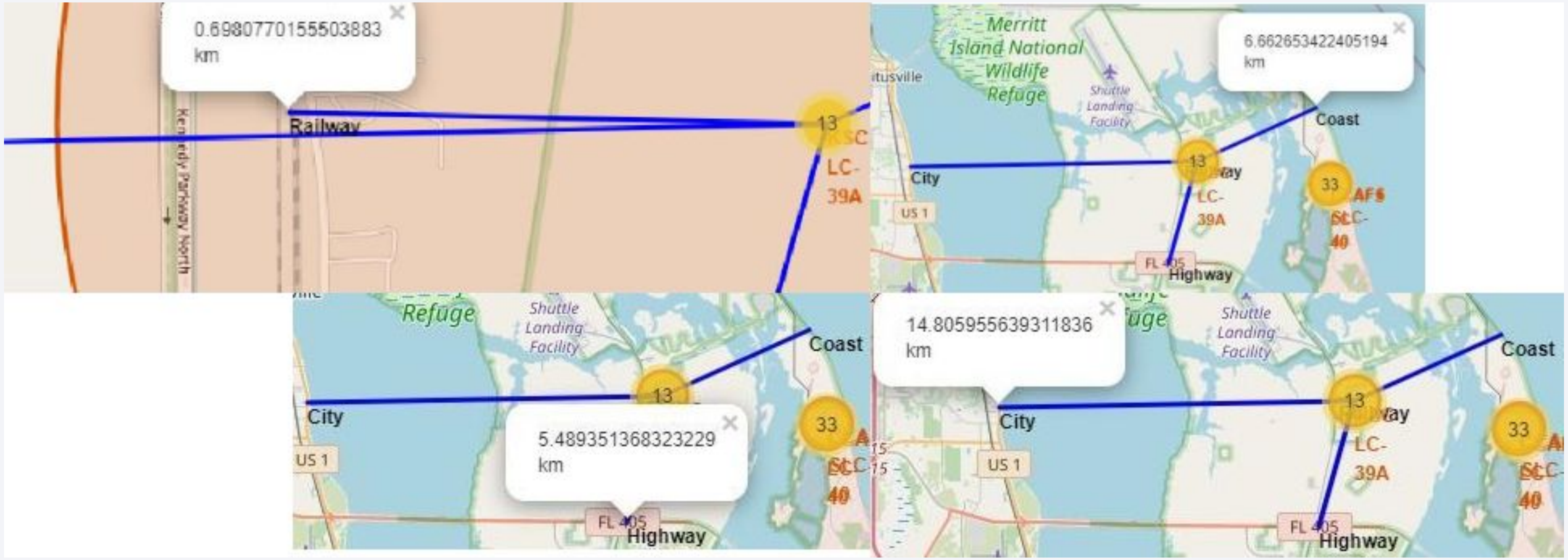
The left map shows all launch sites relative US map. The right map shows the two Florida launch sites since they are very close to each other. All launch sites are near the ocean.

<Folium Map Screenshot 2>



Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed
landing (red icon). In this example VAFB SLC-4E shows 4 successful landings and 6 failed landings.

<Folium Map Screenshot 3>

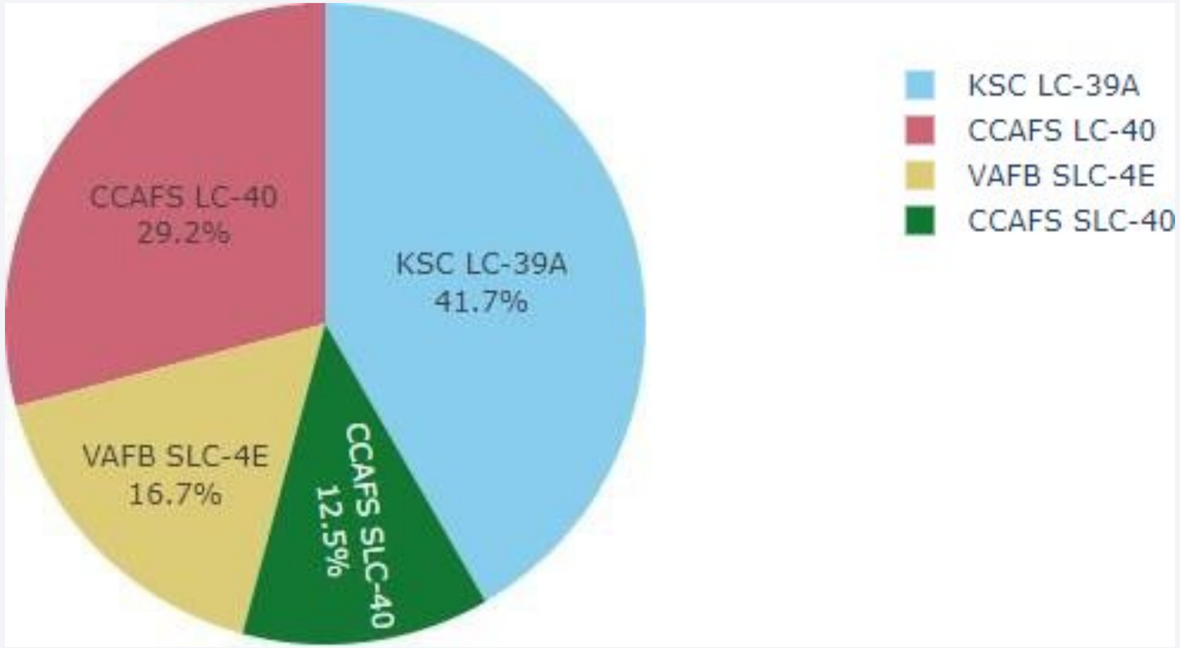


Using KSC LC-39A as an example, launch sites are very close to railways for large part and supply transportation. Launch sites are close to highways for human and supply transport. Launch sites are also close to coasts and relatively far from cities so that launch failures can land in the sea to avoid rockets falling on densely populated areas.

Section 4

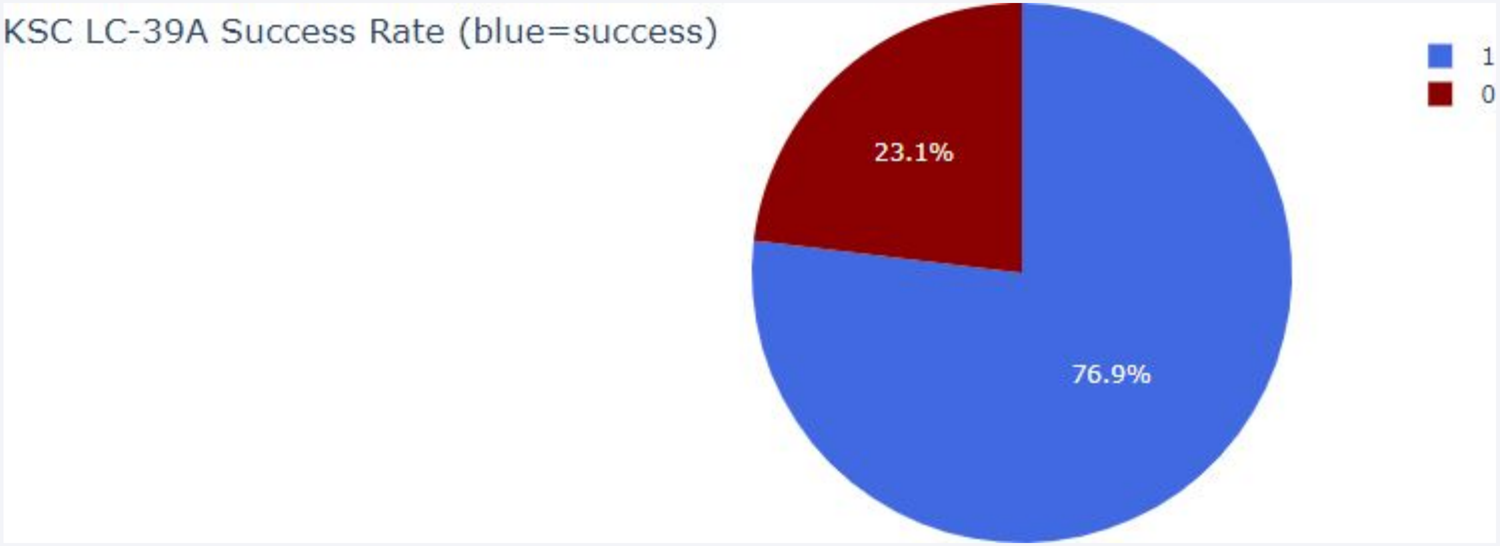
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>



- This is the distribution of successful landings across all launch sites. CCAFS LC-40 is the old name of CCAFS SLC-40 so CCAFS and KSC have the same amount of successful landings.
- VAFB has the smallest share of successful landings. This is first of all it was less used launching site and second because it was use mostly in the early years when Space X didn’t have good success record.

<Dashboard Screenshot 2>



- KSC LC-39A has the highest success rate with 10 successful landings and 3 failed landings.

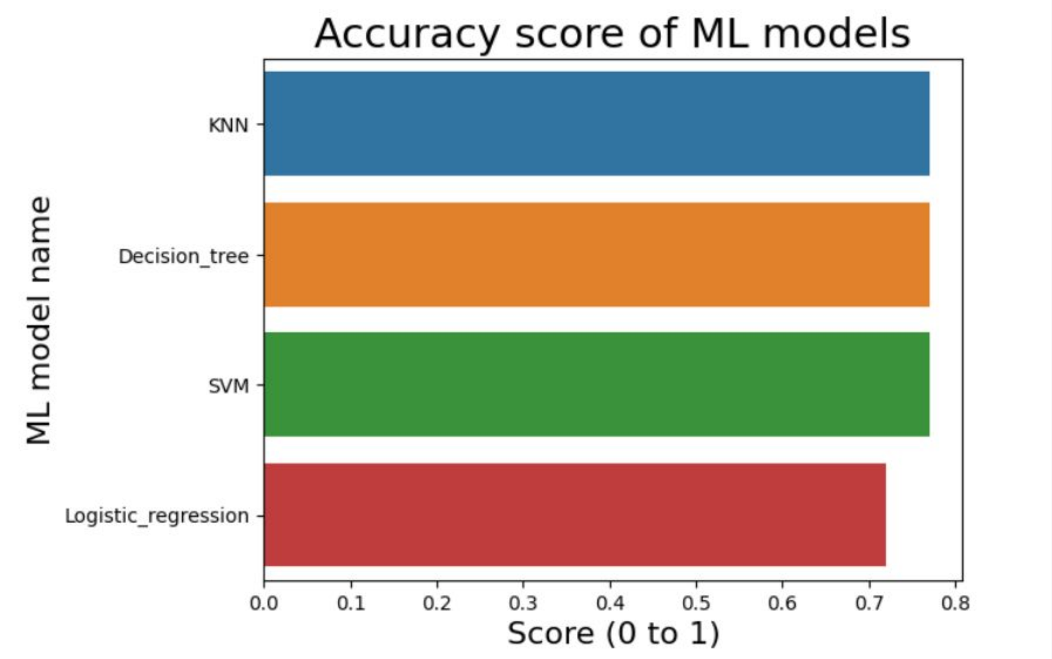
<Dashboard Screenshot 3>

- 1 indicates successful landing and 0 failed landing.
- The booster version category is color coded

Section 5

Predictive Analysis (Classification)

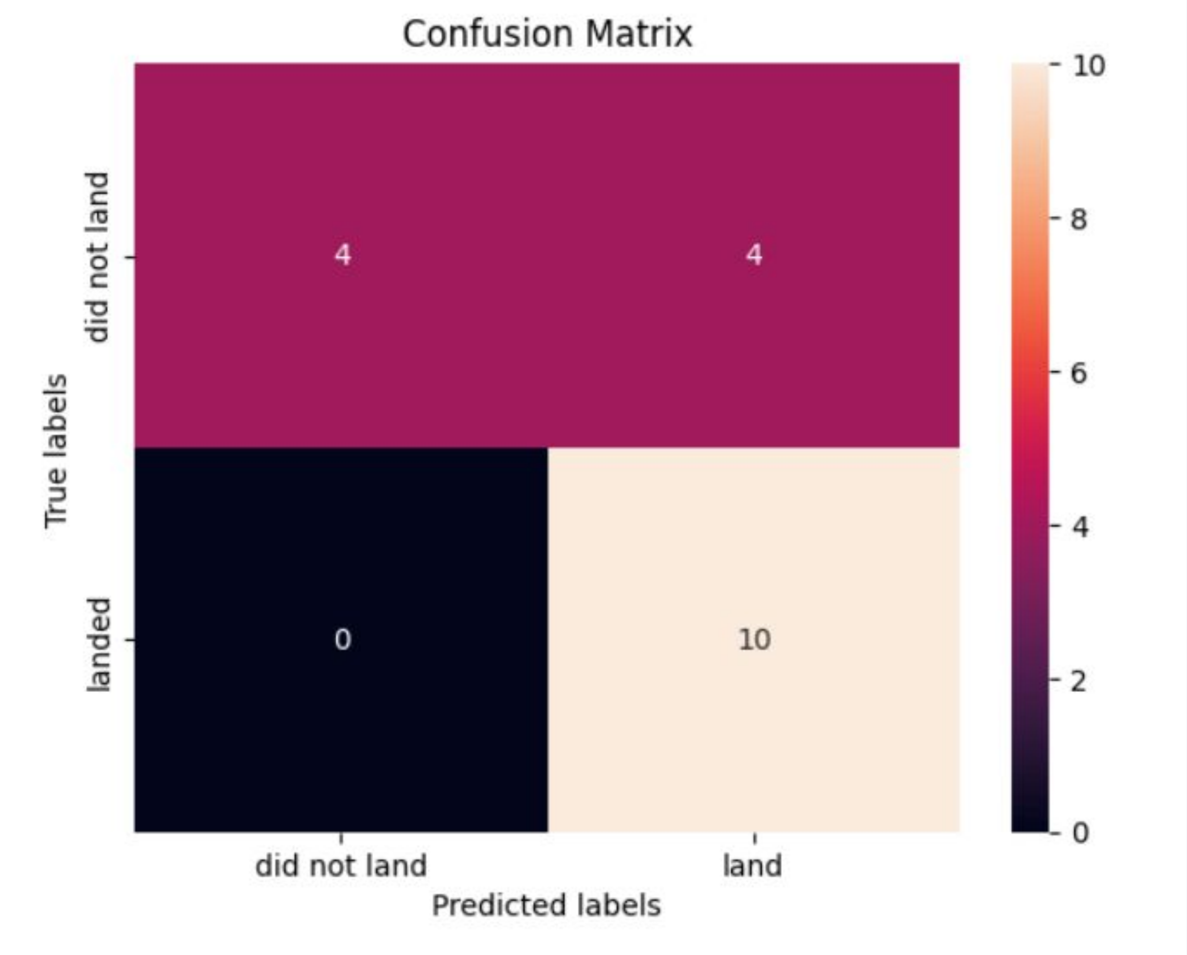
Classification Accuracy



- All models (except logistic regression) had virtually the same accuracy on the test set at 77% accuracy.
- It should be noted that test size is small at only sample size of 18.
- This can cause large variance in accuracy results, such as those in Decision Tree Classifier model in repeated runs.
- We likely need more data to determine the best model.

Confusion Matrix

- All models performed the same for the test set (except logistic regression), thus the confusion matrix is the same across all models. The models predicted 10 successful landings when the true label was successful landing.
- The models predicted 4 unsuccessful landings when the true label was unsuccessful landing.
- The models predicted 4 successful landings when the true label was unsuccessful landings (false positives).
- Our models over predict successful landings.



Conclusions

- It is clear that each year on average the success rates of the launches/landings is getting better
- Allon Mask of SpaceY can use our model to predict with ~77% accuracy whether a launch will have a successful Stage 1 landing before launch to determine whether the launch should be made or not
- If possible more data should be collected to better determine the best machine learning model and improve accuracy
- It’s clear that time was the strongest factor in Space X’s landing successes. Thus, experts from Space X with the right domain knowledge could prove imperative to understanding what makes a launch/landing successful or not.

Appendix

GitHub repository:
https://github.com/Gandalfdore/Data-Science-Projects/tree/main/SpaceX-CapstoneProject

Special Thanks to All Instructors:
https://www.coursera.org/professional-certificates/ibm-data-science?#instructors

Thank you!