Winning Space Race with Data Science

<Name> Isshin Mokushi
<Date> 2026-02-23

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- This project analyzes SpaceX Falcon 9 launch data to understand the factors influencing landing success and to build predictive models for landing outcomes.
- Data was collected through the SpaceX REST API and web scraping of the Falcon 9 Wikipedia Page.
- After data wrangling and feature engineering, EDA was conducted using Python visualization.
- Interactive analytics were developed using Folium and Plotly Dash to explore geospactial and payload-related patterns.
- Four classification models, Logistic Regression, SVM, Decision Tree, and KNN were tuned using GridSearechCV.
- The best-performing model was KNN, and this was selected based on test accuracy and confusion matrix evaluation.

Introduction

Project Background

SpaceX has revolutionized space transportation by reusing Falcon 9 boosters. Booster recovery is essential for reducing launch costs and improving operational efficiency.

Research Questions

- What factors influence Falcon 9 landing success?
- Can landing outcomes be predicted using payload, orbit, launch site, and booster characteristics?
- Which classification model provides the highest predictive accuracy?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
- Perform data wrangling
- 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 Collection – SpaceX API

- Extract rocket launch data from SpaceX API regarding:
  - Rocket
  - Payloads
  - Launchpad
  - Cores
- Convert data into Pandas df
- Filter data to only extract Falcon 9 launches' information.
- Tidy data via replacing missing values of "payloadMass" with own mean.

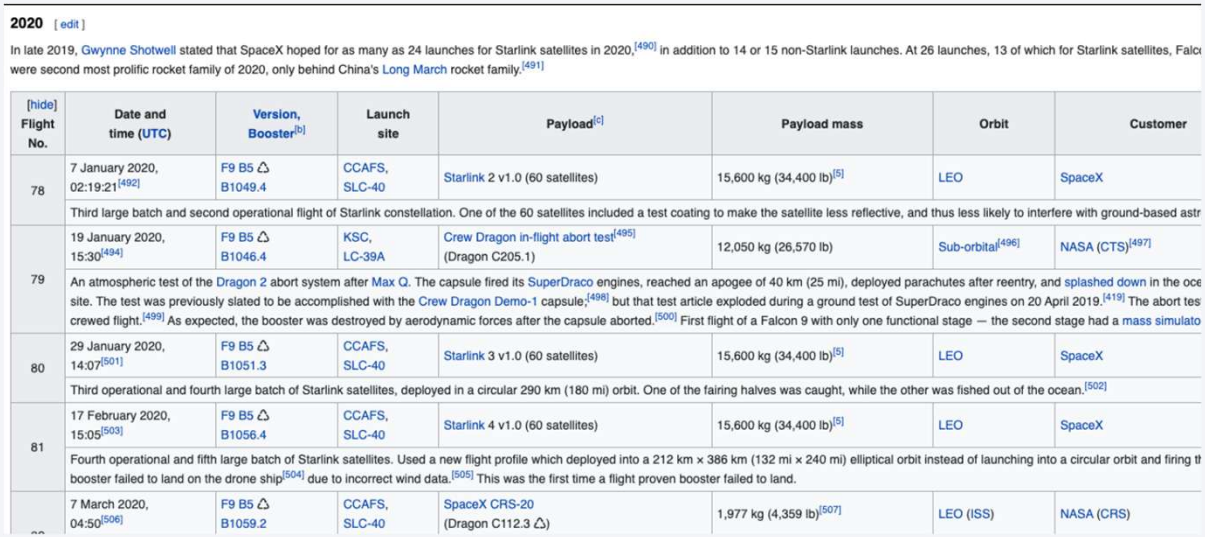
```
data_falcon9.loc[:,'FlightNumber'] = list(range(1, data_falcon9.shape[0]+1))
data_falcon9
```

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	R
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	

URL:edX_finalexam/1_jupyter-labs-spacex-data-collection-api.ipynb at main · UPAUPA-DB/edX_finalexam

Data Collection - Scraping

- Request to wiki page of Falcon 9 launch using its URL
- Extract chart information via HTML tag
- Convert HTML table to pandas df



2020 [edit]

In late 2019, Gwynne Shotwell stated that SpaceX hoped for as many as 24 launches for Starlink satellites in 2020,[490] in addition to 14 or 15 non-Starlink launches. At 26 launches, 13 of which for Starlink satellites, Falc were second most prolific rocket family of 2020, only behind China's Long March rocket family.[491]

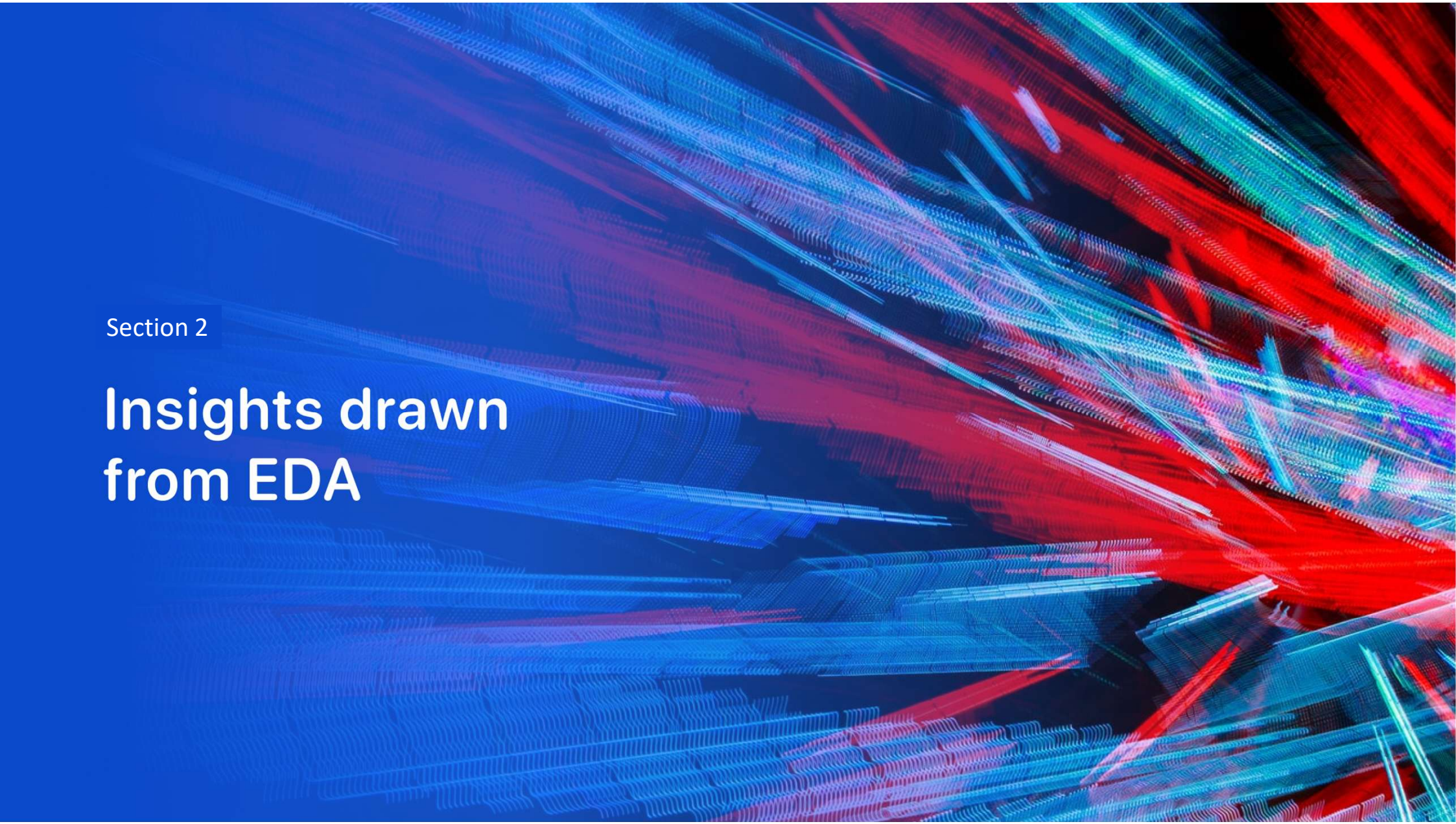
[hide] Flight No.	Date and time (UTC)	Version, Booster[b]	Launch site	Payload[c]	Payload mass	Orbit	Customer
78	7 January 2020, 02:19:21[492]	F9 B5 △ B1049.4	CCAFS, SLC-40	Starlink 2 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX
	Third large batch and second operational flight of Starlink constellation. One of the 60 satellites included a test coating to make the satellite less reflective, and thus less likely to interfere with ground-based astr						
79	19 January 2020, 15:30[494]	F9 B5 △ B1046.4	KSC, LC-39A	Crew Dragon in-flight abort test[495] (Dragon C205.1)	12,050 kg (26,570 lb)	Sub-orbital[496]	NASA (CTS)[497]
	An atmospheric test of the Dragon 2 abort system after Max Q. The capsule fired its SuperDraco engines, reached an apogee of 40 km (25 mi), deployed parachutes after reentry, and splashed down in the oce site. The test was previously slated to be accomplished with the Crew Dragon Demo-1 capsule;[498] but that test article exploded during a ground test of SuperDraco engines on 20 April 2019.[419] The abort tes crewed flight.[499] As expected, the booster was destroyed by aerodynamic forces after the capsule aborted.[500] First flight of a Falcon 9 with only one functional stage — the second stage had a mass simulato						
80	29 January 2020, 14:07[501]	F9 B5 △ B1051.3	CCAFS, SLC-40	Starlink 3 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX
	Third operational and fourth large batch of Starlink satellites, deployed in a circular 290 km (180 mi) orbit. One of the fairing halves was caught, while the other was fished out of the ocean.[502]						
81	17 February 2020, 15:05[503]	F9 B5 △ B1056.4	CCAFS, SLC-40	Starlink 4 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX
	Fourth operational and fifth large batch of Starlink satellites. Used a new flight profile which deployed into a 212 km × 386 km (132 mi × 240 mi) elliptical orbit instead of launching into a circular orbit and firing th booster failed to land on the drone ship[504] due to incorrect wind data.[505] This was the first time a flight proven booster failed to land.						
	7 March 2020, 04:50[506]	F9 B5 △ B1059.2	CCAFS, SLC-40	SpaceX CRS-20 (Dragon C112.3 △)	1,977 kg (4,359 lb)[507]	LEO (ISS)	NASA (CRS)

URL:edX_finalexam/2_jupyter-labs-webscraping.ipynb at main · UPAUPA-DB/edX_finalexam

Data Wrangling

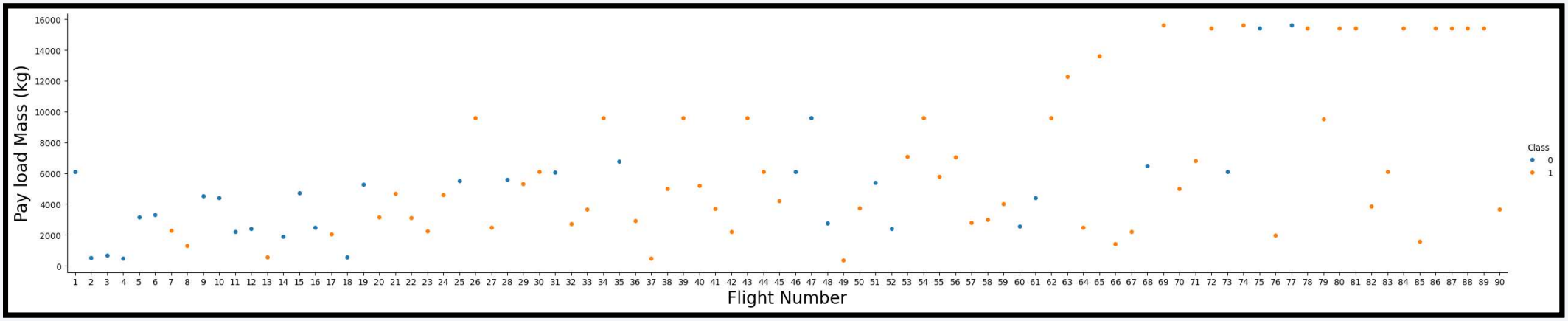
- Conduct EDA:
  - Calculate number of launches by site
  - Calculate the number and occurrence of each orbit
  - Calculate the number and occurrence of mission outcome of the orbits
- Create a landing outcome label
  - Definition of bad outcomes : {'False ASDS', 'False Ocean', 'False RTLS', 'None ASDS', 'None None'}
  - If outcome is included in the above list, then mutating "Landing_class" column with 0 (unsuccessful landing).
  - If outcome is NOT included in the above list, then mutating "Landing_class" column with 1 (successful landing).

URL:edX_finalexam/3_labs-jupyter-spacex-data wrangling_jupyterlite.ipynb at main · UPAUPA-DB/edX_finalexam


Section 2
Insights drawn from EDA

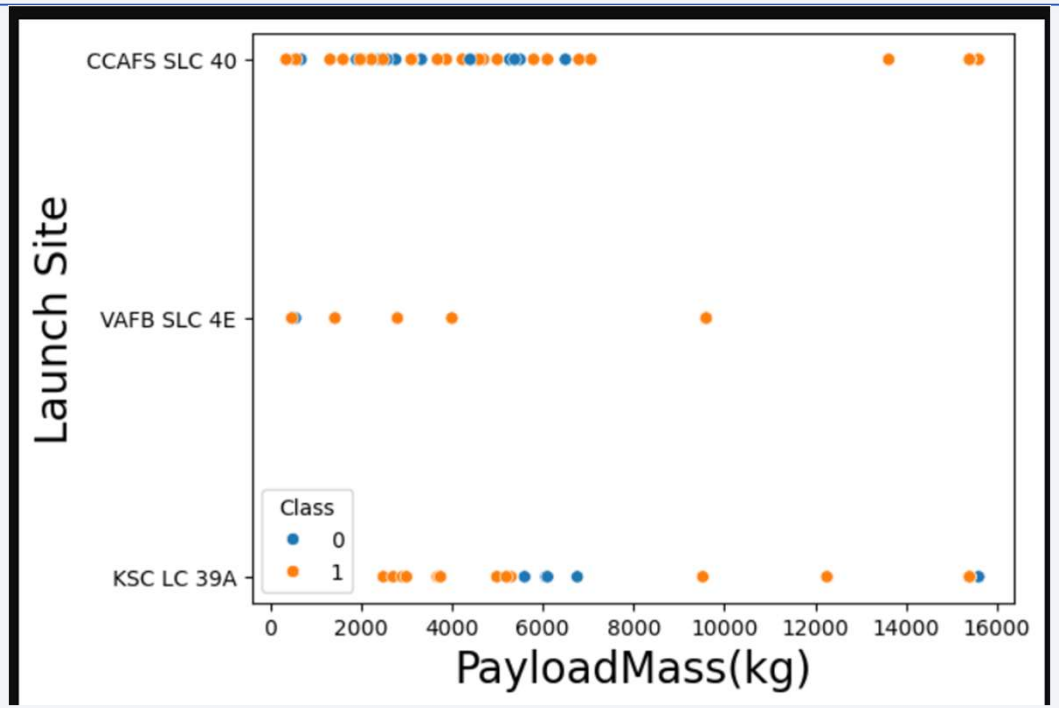
EDA with Data Visualization (Flight Number vs Payload Mass)

- First, see how the “FlightNumber” (indicating the continuous launch attempts.) and “Payload” variables would affect the launch outcome.



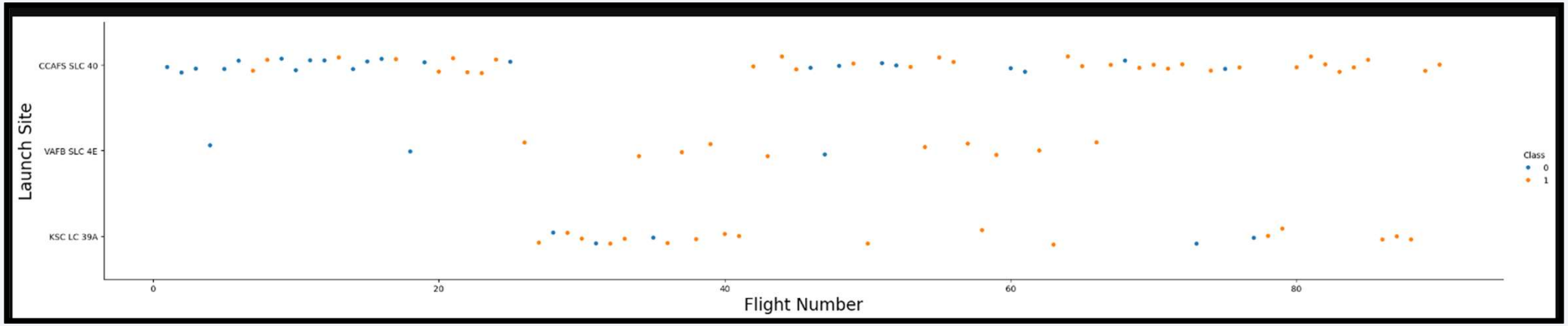
- We see that as the flight number increases, the first stage is more likely to land successfully. The payload mass also appears to be a factor; even with more massive payloads, the first stage often returns successfully.

EDA with Data Visualization (Payload Mass vs Launch Site)



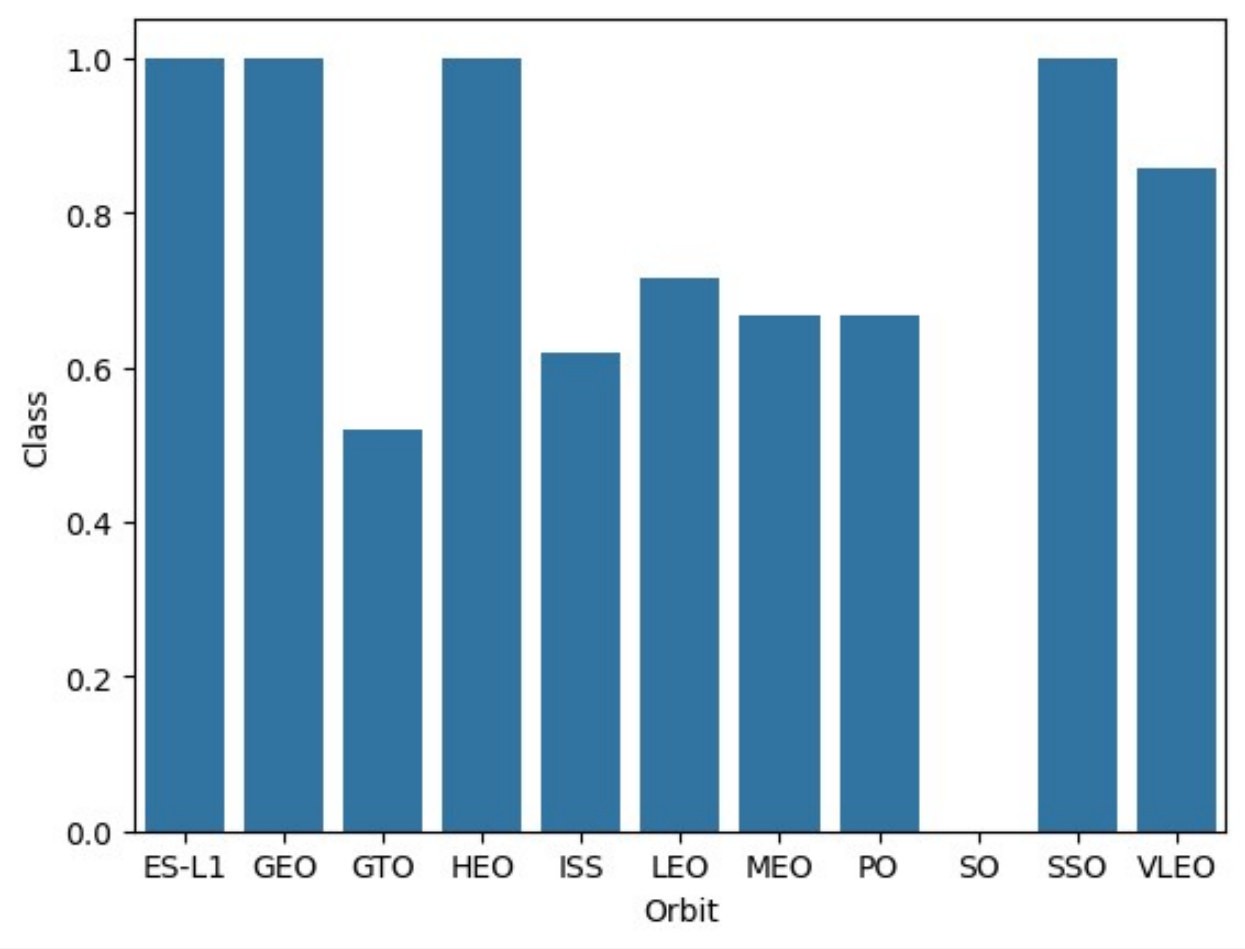
- We see that there are no rockets launched for heavy payload mass(greater than 10000) in VAFB-SLC and it appears that the scale of payload mass does not affect the successful rate of launch.

EDA with Data Visualization (Flight Number vs Launch Site)

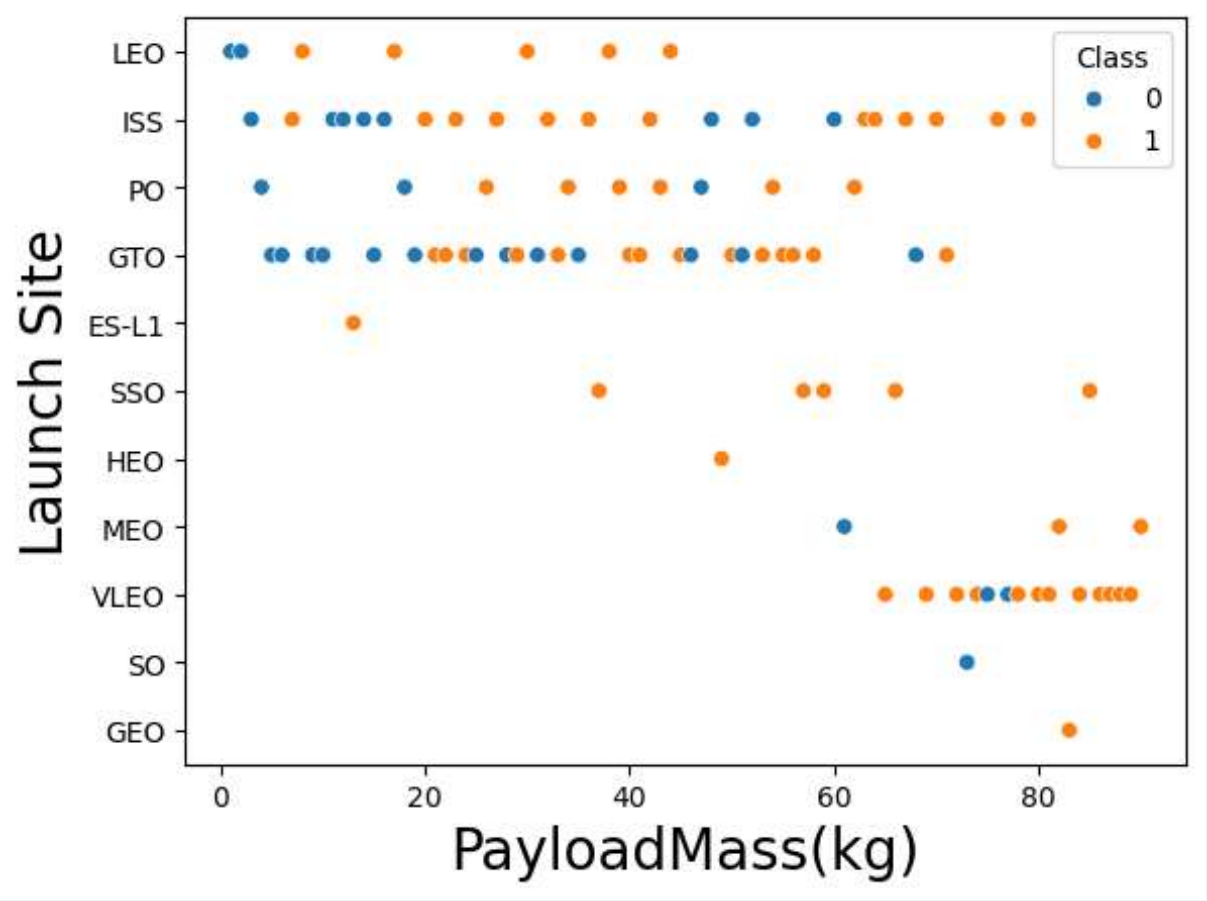


- We see more increasing flight number, the successful rate of launching become better in every launch site.

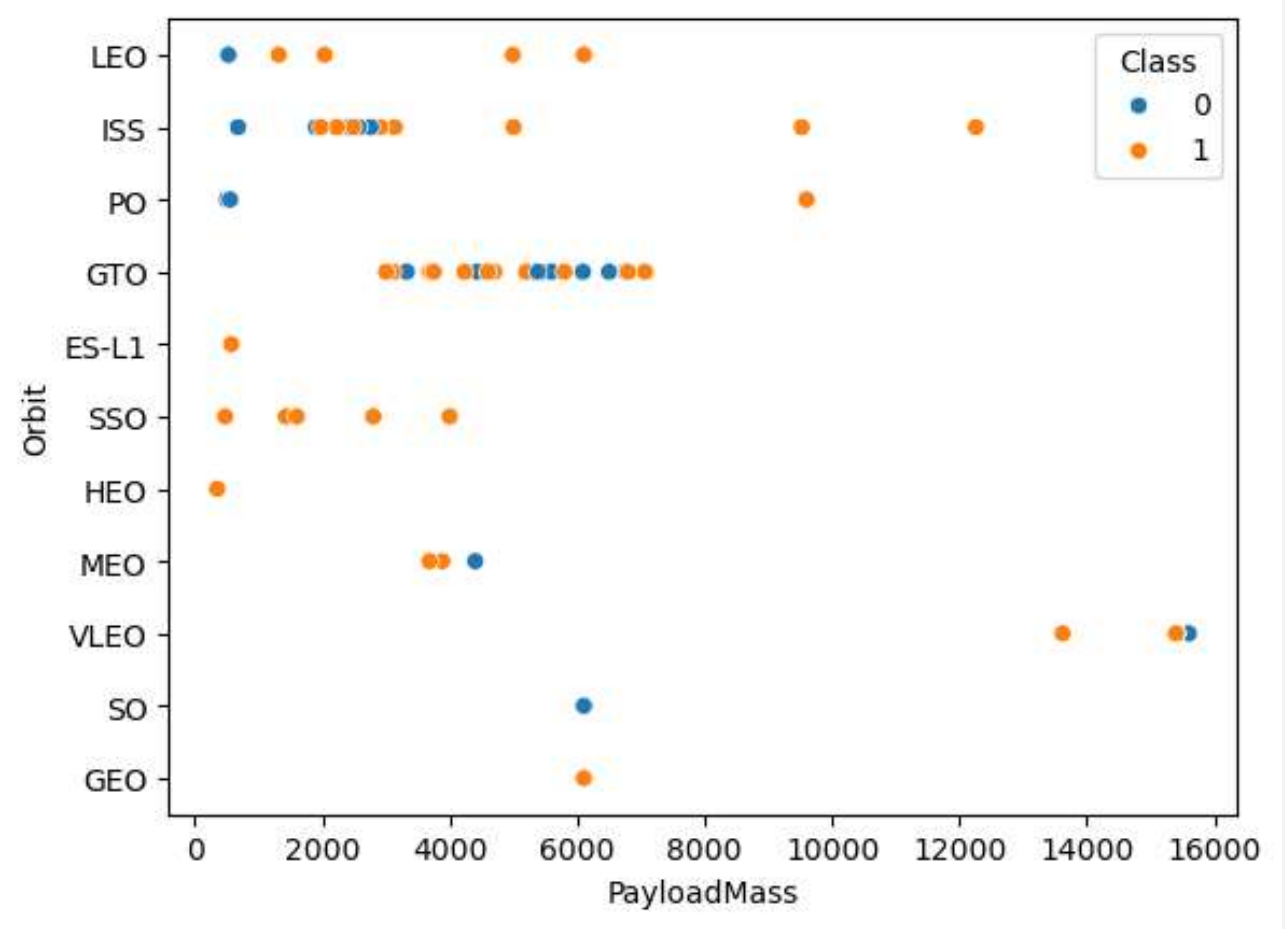
EDA with Data Visualization (Success Rate by Orbit type)



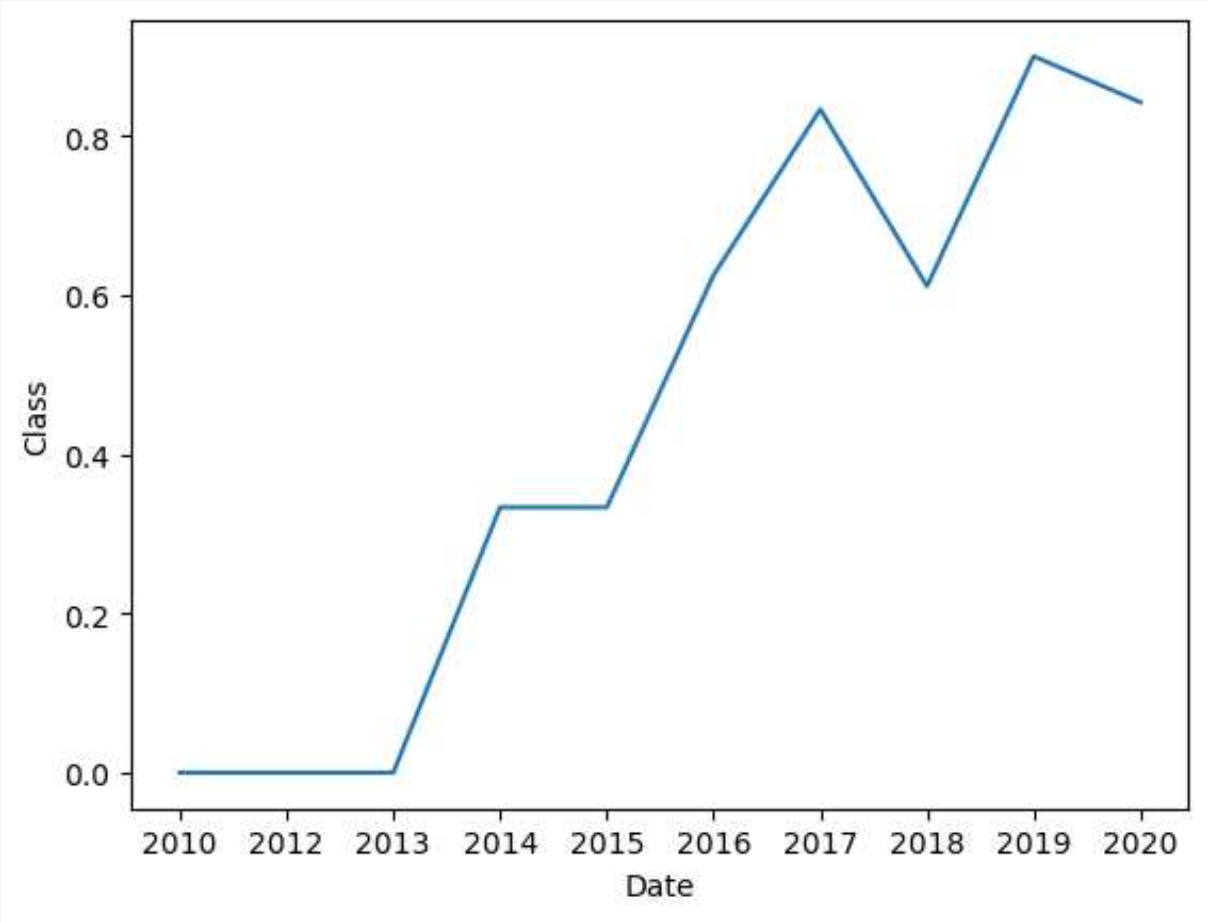
EDA with Data Visualization (Flight Number vs Orbit)



EDA with Data Visualization (Payload Mass vs Orbit)



EDA with Data Visualization (Launch success yearly trend)



EDA with Data Visualization (Primary Findings)

- The probability of succeeding in launch will be improved by the increasing challenges – this means accumulating experiences.
- The probability of succeeding in launch had been increasing through 2013 and 2020.
- Based on several visualization EDA, we can assume that the following variables will have an effect on the success rate:
  - Orbits
  - Launch Site
  - Flight Number

URL:edX_finalexam/7_edadataviz.ipynb at main · UPAUPA-DB/edX_finalexam

EDA with SQL Outcomes

※ I got an unknown error regarding SQL, so all the results were generated via Python code instead of SQL

1. Names of the launch sites:

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

2. Head of records where launch sites begin with “KSC-”:

	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	# PAYLOAD_MASS__KG_	Orbit
29	2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)
30	2017-03-16	6:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO
31	2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO
32	2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO
33	2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO

EDA with SQL Outcomes

※ I got an unknown error regarding SQL, so all the results were generated via Python code instead of SQL

3. Total payloads mass carried by boosters launched by NASA (CRS):

    45,596 kg

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

    2,928.4 kg

5. Date of the first landing success in drone ship is:

    2016-04-08

EDA with SQL Outcomes

※ I got an unknown error regarding SQL, so all the results were generated via Python code instead of SQL

6. Boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000:

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

7. Total number of successful and failure mission outcomes:

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

EDA with SQL Outcomes

※ I got an unknown error regarding SQL, so all the results were generated via Python code instead of SQL

8. Booster versions that have carried the maximum payload mass:

Booster_Version	# Booster_Version
F9 B5 B1048.4	1
F9 B5 B1048.5	1
F9 B5 B1049.4	1
F9 B5 B1049.5	1
F9 B5 B1049.7	1
F9 B5 B1051.3	1
F9 B5 B1051.4	1
F9 B5 B1051.6	1
F9 B5 B1056.4	1
F9 B5 B1058.3	1

Booster_Version	# Booster_Version
F9 B5 B1060.2	1
F9 B5 B1060.3	1

URL:edX_finalexam/4_jupyter-labs-eda-sql-edx-sqllite.ipynb at main · UPAUPA-DB/edX_finalexam

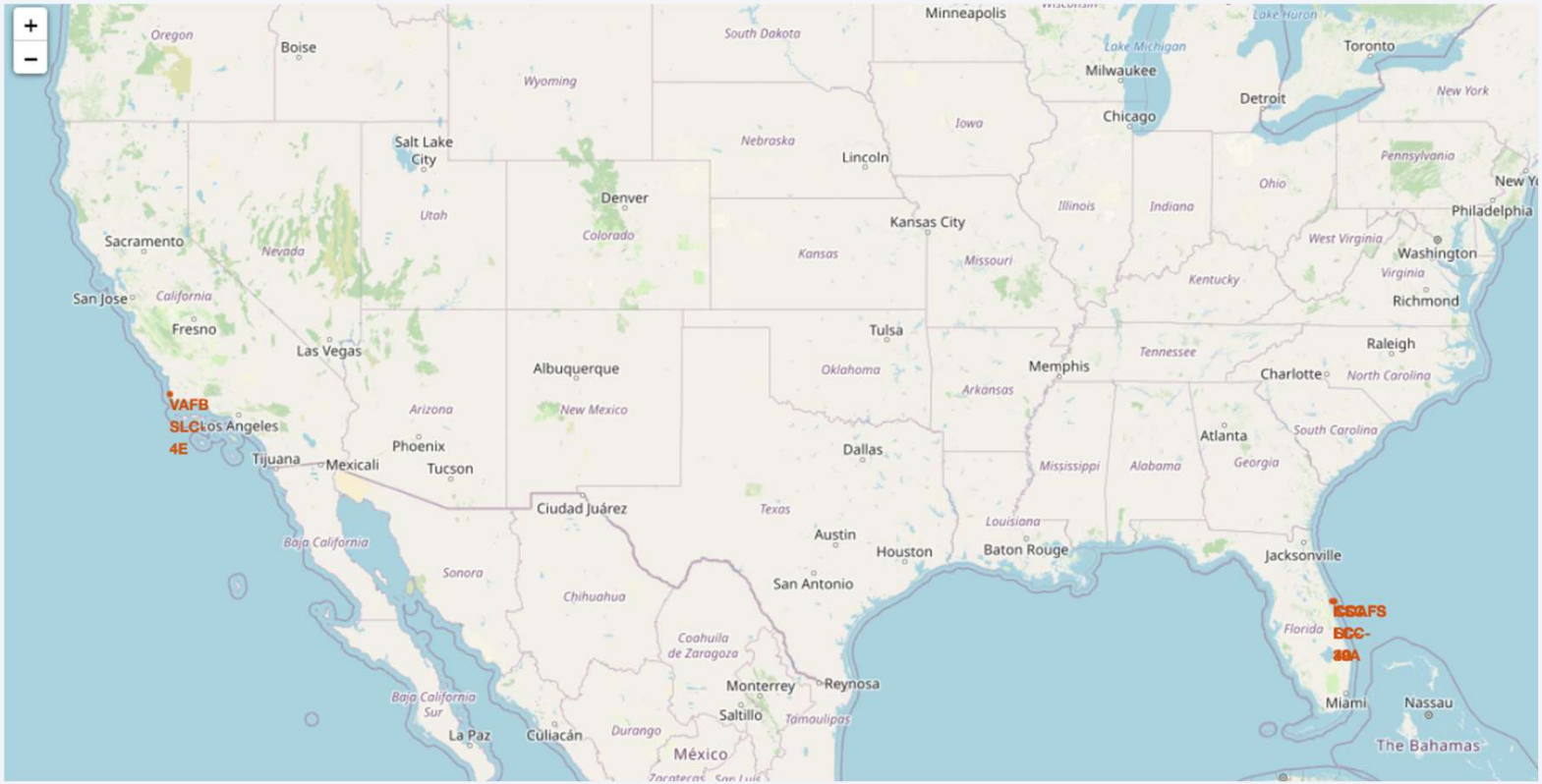

Section 3
Launch Sites
Proximities Analysis

Build an Interactive Map with Folium

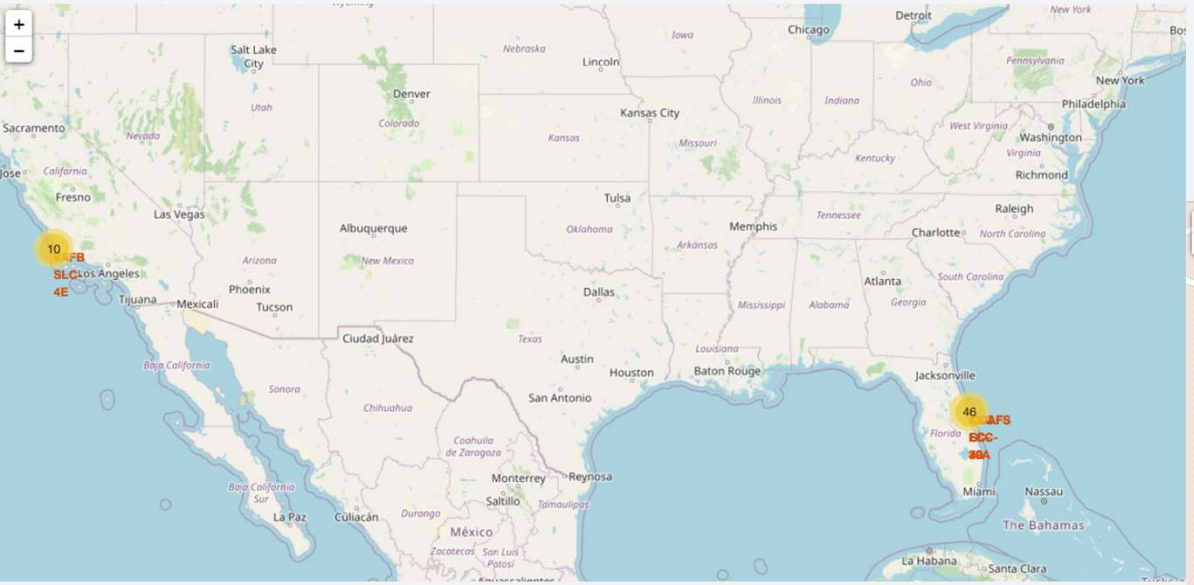
- Added markers for all launch sites.
- Added color-coded makers for landing outcomes.
- Added circles and distance lines to visualize proximity to coastlines, highways, and railways.
- Finally, we found that:
  - Launch sites are near of highway and rail since utilizing the transportation.
  - Launch sites are near of coastline and equator to utilize the rotation of the earth.
  - Launch sites are distant from cities for safety reason,

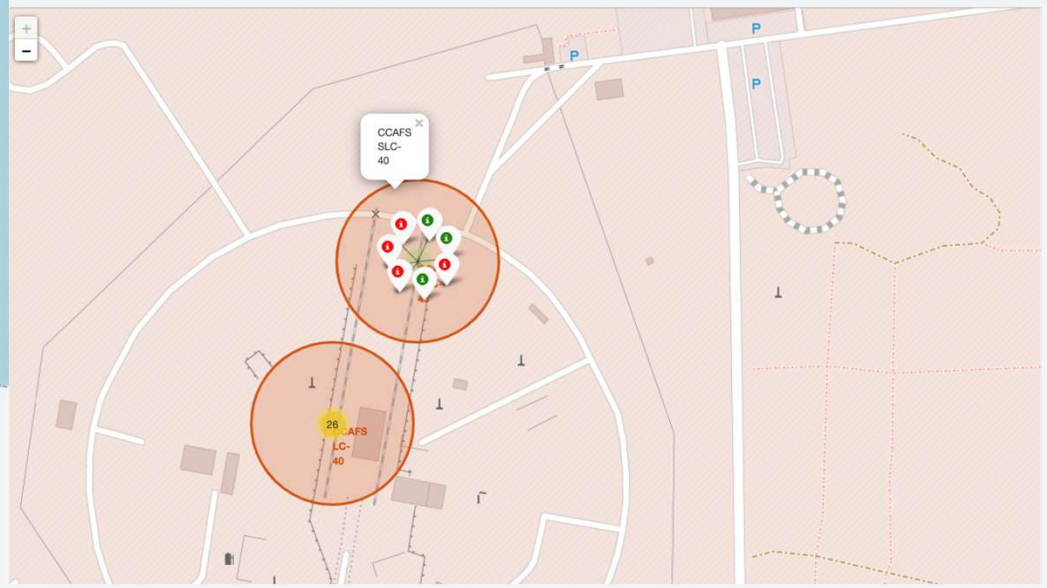
URL:edX_finalexam/6_labs_jupyter_launch_site_location.ipynb at main · UPAUPA-DB/edX_finalexam

<Folium Map Screenshot 1> Launch Sites

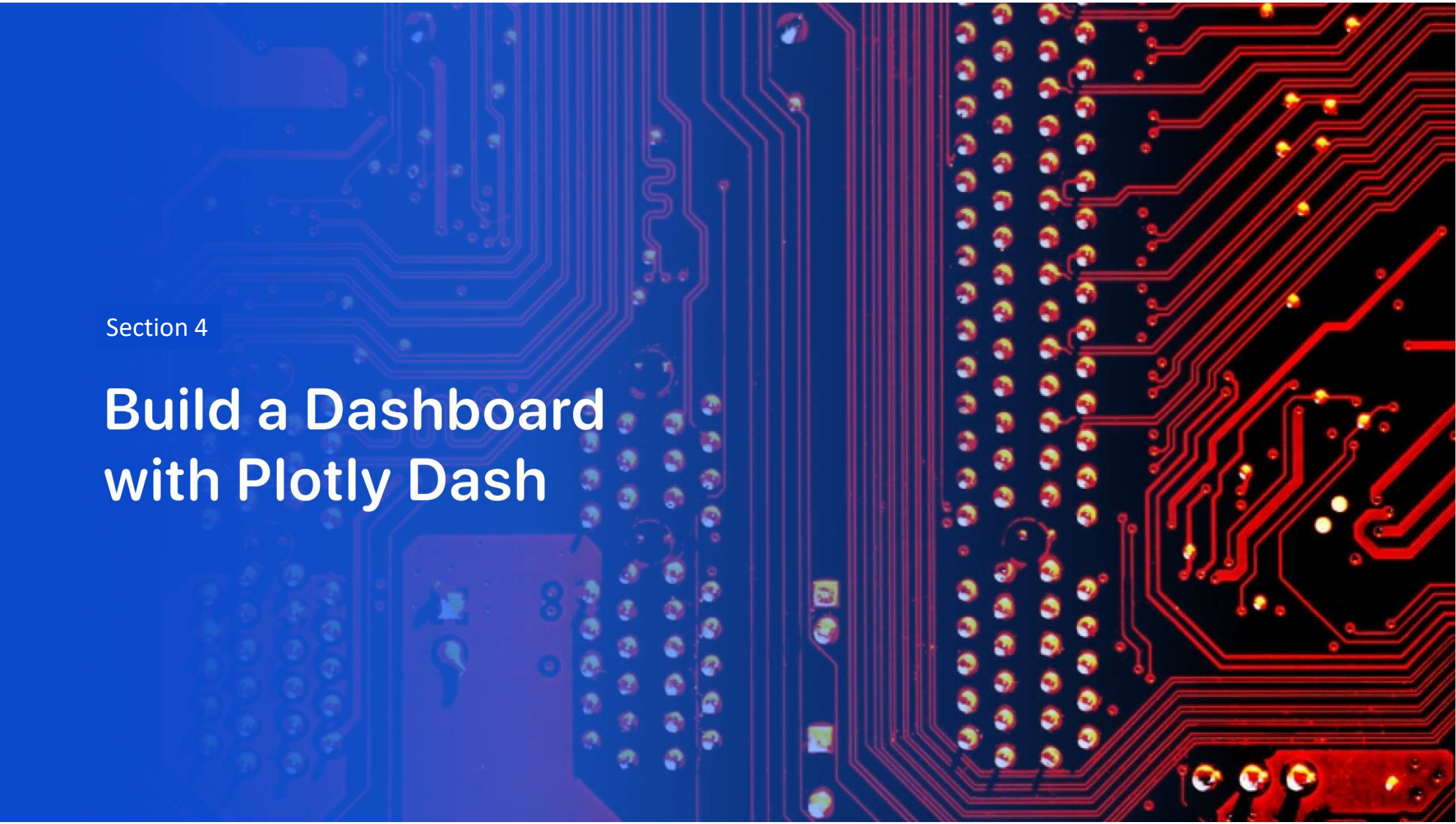


<Folium Map Screenshot 2> Successful and Failed Launches





<Folium Map Screenshot 3>

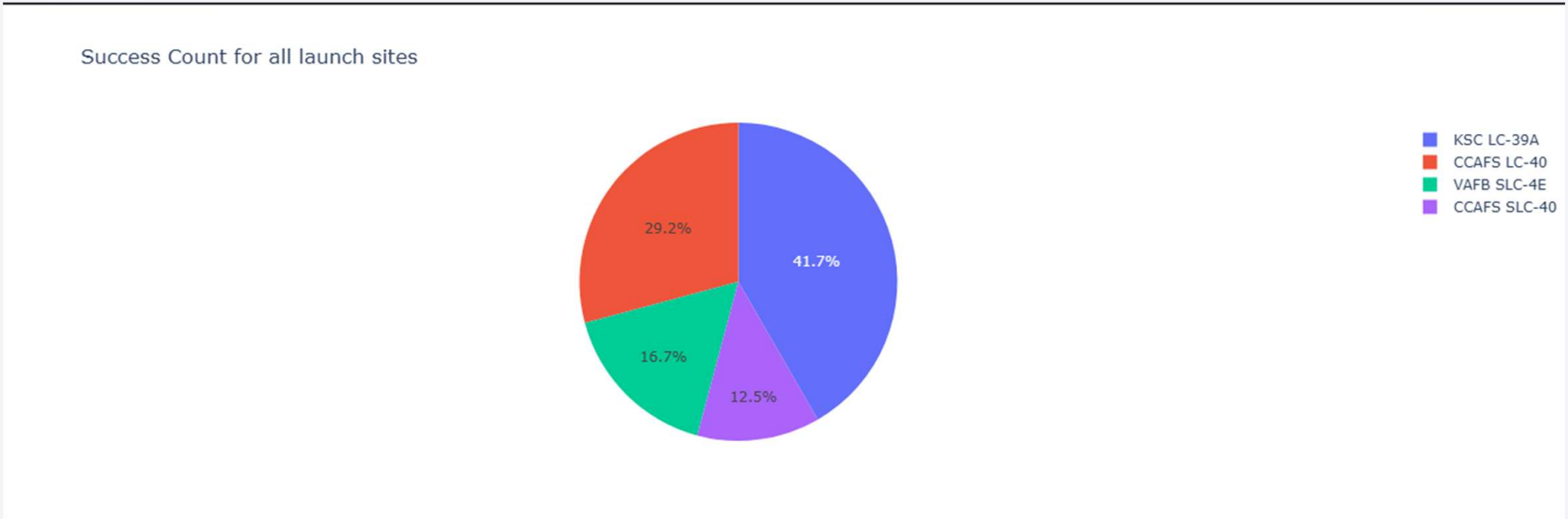

Section 4
Build a Dashboard with Plotly Dash

Build a Dashboard with Plotly Dash

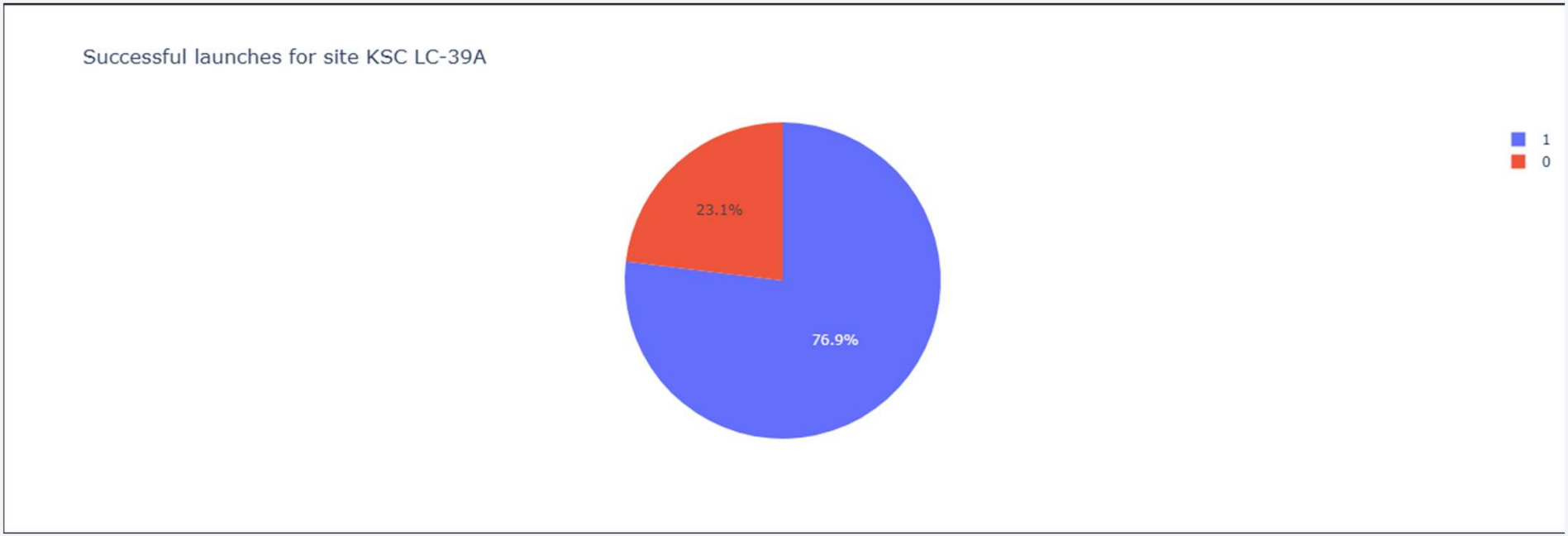
- Dashboard components:
  - Pie Chart : Launch success count by site
  - Pie Chart : Success ratio for selected site
  - Scatter plot : Payload vs Launch Outcome with range slider
  - Interactive filters for site and payload range
- Dashboard analytics finding:
  - KSC occupies approximately 40% of success cases in total.
  - Payload range 0 to 2,000 and 6,000 to 9,000 have the lowest success rate

URL:UPAUPA-DB/edX_finalexam

<Dashboard Screenshot 1> Success Count for all launch sites



<Dashboard Screenshot 2> Successful launches for site KSC LC-39A



<Dashboard Screenshot 3> Success count on Payload mass for all sites

Section 5

Predictive Analysis (Classification)

Predictive Analysis (Classification)

- Standardized features and split into train and test sets
- Trained via four models:
  - Logistic Regression
  - SVM
  - Decision Tree
  - KNN
- Used GridSearchCV(cv=10) to tune hyperparameters
- Evaluated models using test accuracy
- Selected the best model and analyzed its confusion matrix

Classification Accuracy

```
print("tuned hpyerparameters :(best parameters) ",logreg_cv.best_params_)
print("accuracy :",logreg_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
accuracy : 0.8464285714285713
```

```
print("tuned hpyerparameters :(best parameters) ",svm_cv.best_params_)
print("accuracy :",svm_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'C': 1.0, 'gamma': 0.03162277660168379, 'kerne
l': 'sigmoid'}
accuracy : 0.8482142857142856
```

```
print("tuned hpyerparameters :(best parameters) ",tree_cv.best_params_)
print("accuracy :",tree_cv.best_score_)
```

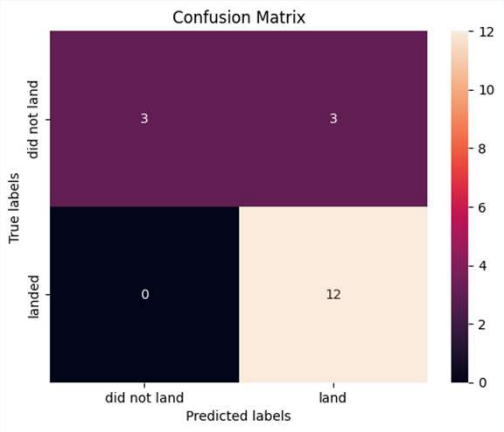
```
tuned hpyerparameters :(best parameters)  {'criterion': 'entropy', 'max_depth': 16, 'max_f
eatures': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 5, 'splitter': 'random'}
accuracy : 0.8892857142857145
```

```
print("tuned hpyerparameters :(best parameters) ",knn_cv.best_params_)
print("accuracy :",knn_cv.best_score_)
```

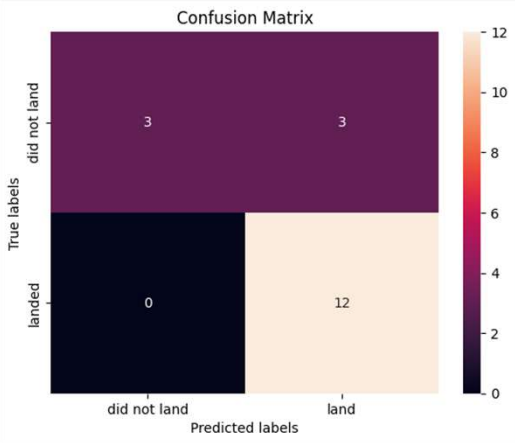
```
tuned hpyerparameters :(best parameters)  {'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}
accuracy : 0.8482142857142858
```

Confusion Matrix

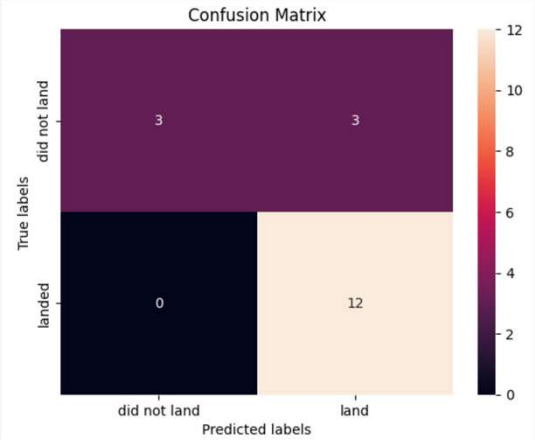
Logistic Regression



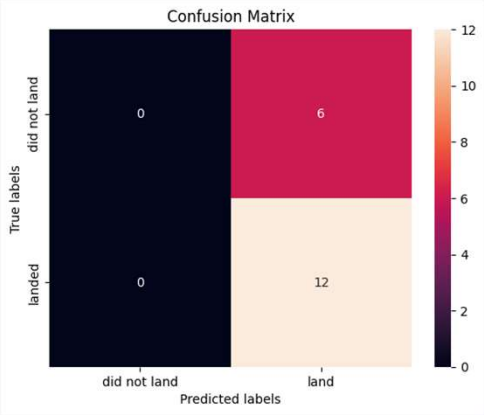
KNN



SVM



Decision Tree



Results

- EDA revealed strong relationships between payload mass, orbit type, and landing success.
- Folium maps highlighted geographic patterns and launch site characteristics.
- Dash dashboard enabled interactive exploration of payload and success trends.
- The best-performing model achieved the highest accuracy (e.g., Logistic Regression).
- Confusion matrix showed strong predictive performance with minimal misclassification.

URL:UPAUPA-DB/edX_finalexam

Conclusions

- Falcon 9 landing success is influenced by payload mass, orbit type, launch site, and booster experience.
- Machine learning models can effectively predict landing outcomes.
- The best model provides actionable insights for mission planning and risk reduction.
- This analytical framework can be extended to future SpaceX missions and other launch vehicles.

Thank you!