



IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Hanim Basarudin
December 4th 2024



Outline

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

Executive Summary

- Summary of methodologies
 - Data collection with API & Web Scraping
 - Data wrangling
 - Exploratory data analysis (EDA)
 - Interactive visual analytics & dashboard
 - Predictive analysis
- Summary of all results
 - EDA results
 - Interactive analytics screenshots
 - Prediction results

Introduction



Commercial space age

- Emergence of companies making space travel accessible
- Key players: Virgin Galactic, Rocket Lab, Blue Origin & SpaceX



Problem statement

- Not all first stages are successfully recovered
- Influencing factors: payload weight, desired orbit, & customer requirements



SpaceX achievements

- Delivering cargo to the ISS
- Launching Starlink satellite constellation
- Conducting crewed missions to space
- Competitive pricing: Falcon 9 at \$62mil vs. \$165mil by other providers



As a data scientist for SpaceY

- Goal: predict first-stage reuse of Falcon 9 using machine learning
- Analyze first-stage recovery factors leading to launch prediction outcomes

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
 - SpaceX REST API & web scrapping with BeautifulSoup
- Perform data wrangling
 - Attributes processing & binary classification launches (1: success, 0: failure)
- 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 splitting, model development, & model performance evaluation, i.e. prediction accuracy

Data Collection

- Data collection is accessing & gathering useful information from various sources to be used for analysis. In this project, there are 2 ways for collecting data:

1

SpaceX REST API

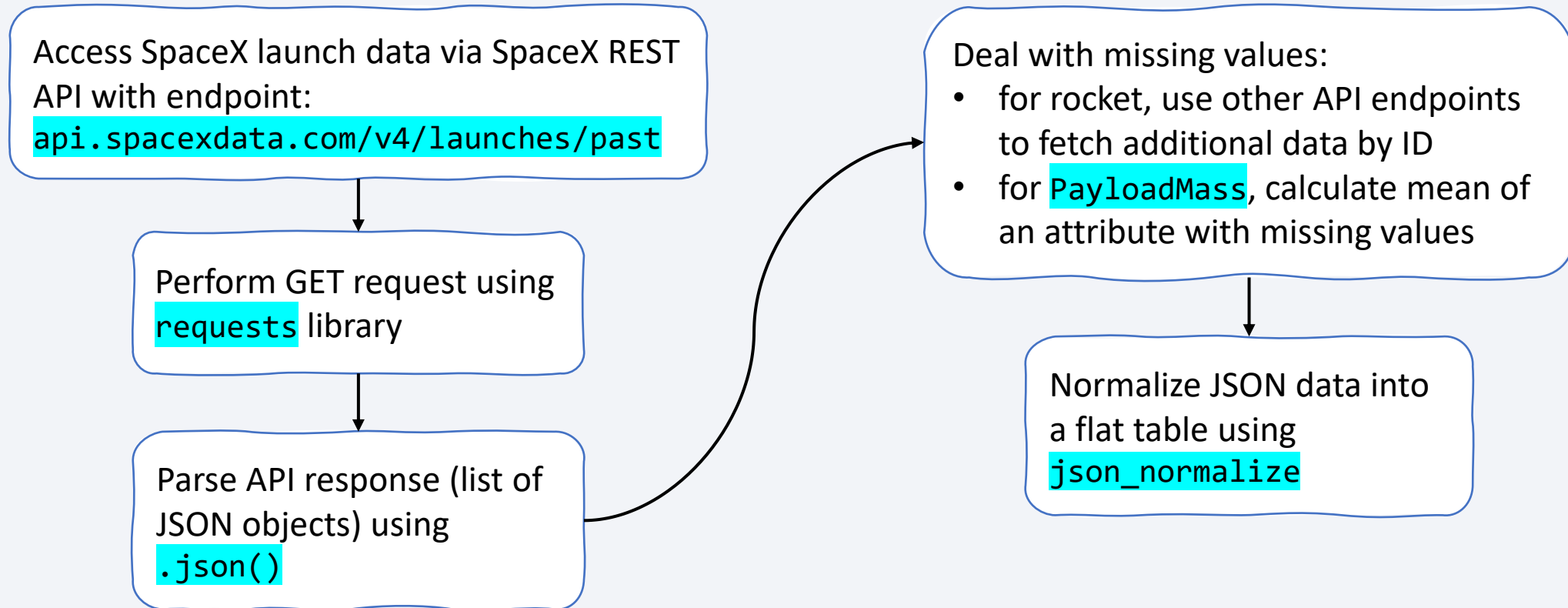
Using endpoints, perform a GET request, parse API response (list of JSON objects), and normalize JSON data into a table.

2

Web scraping

Using Python's BeautifulSoup package, scrape HTML tables from relevant sites, parse, and convert scraped data into Pandas DataFrames.

Data Collection – SpaceX API



<https://github.com/KisMyS/Data-Science-Capstone-IBM-Coursera/blob/main/Ex-1-jupyter-labs-spacex-data-collection-api-v2.ipynb>

Data Collection - Scraping

Use Python's BeautifulSoup package to scrape HTML tables from relevant Wiki pages:

https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Perform GET HTTP request using `requests` library

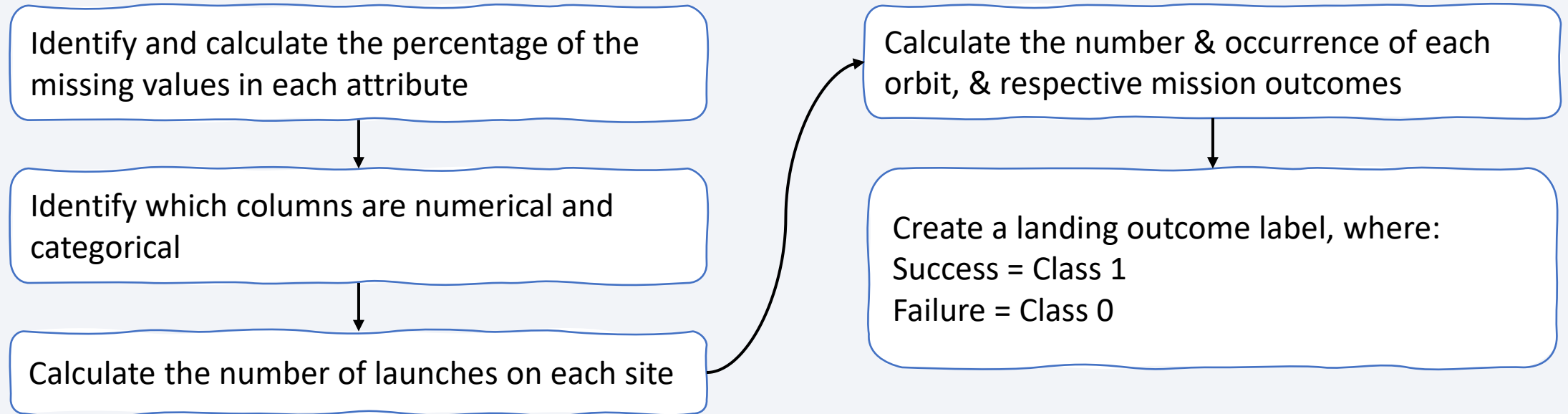
Parse and convert scraped data, including columns & variables names into dataframes.



<https://github.com/KisMyS/Data-Science-Capstone-IBM-Coursera/blob/main/Ex-2-jupyter-labs-webscraping.ipynb>

Data Wrangling

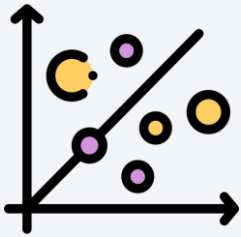
- Data wrangling is cleaning, transforming, & preparing raw data for EDA. In this project, we:



<https://github.com/KisMyS/Data-Science-Capstone-IBM-Coursera/blob/main/Ex-3-labs-jupyter-spacex-Data%20wrangling-v2.ipynb>

EDA with Data Visualization

- To better understand the correlation among attributes, data visualization is performed:



Flight Number vs. Payload
Flight Number vs. Launch Site
Payload vs. Launch Site
Payload vs. Orbit Type
Flight Number vs. Orbit Type



Yearly launch success trend



Success rate by orbit type



<https://github.com/KisMyS/Data-Science-Capstone-IBM-Coursera/blob/main/Ex-5-jupyter-labs-eda-dataviz-v2.ipynb>

EDA with SQL

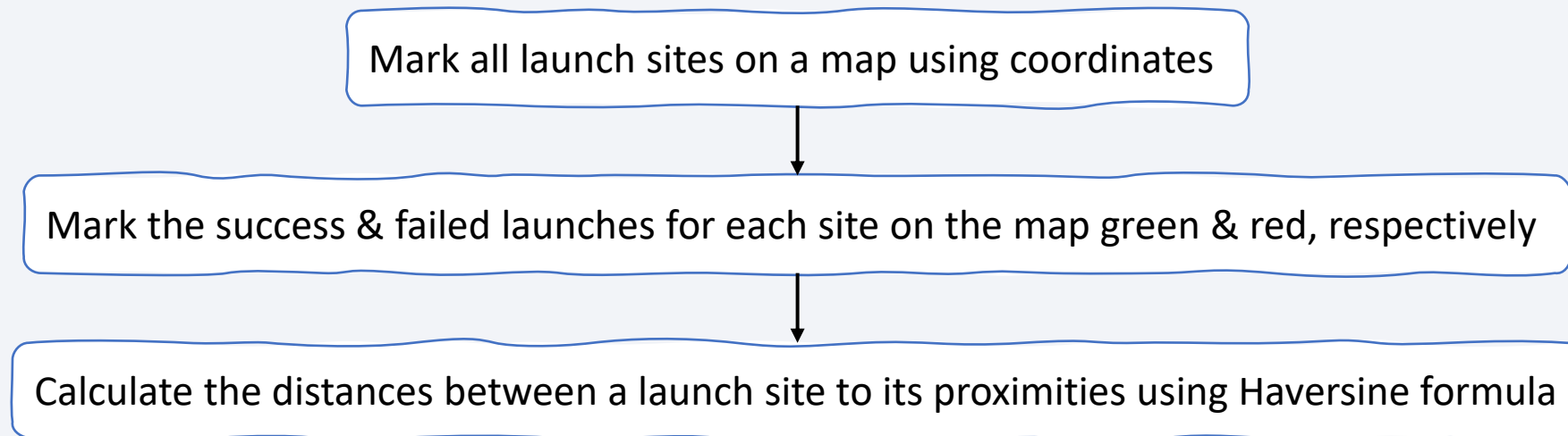
- To get better understanding of the data set, SQL queries are performed:
 - Display names of the unique launch sites in the space mission
 - Display 5 records where launch sites begin with the string 'CCA'
 - Display total payload mass carried by boosters launched by NASA (CRS)
 - Display average payload mass carried by booster version F9 v1.1
 - List date of the first successful landing outcome in ground pad.
 - List booster names having success in drone ship with payload mass between 4000kg & 6000kg
 - List total number of successful & failure mission outcomes
 - List names of booster versions carrying maximum payload mass using subquery
 - List records on month names, failed outcomes in drone ship, booster versions, launch site for 2015.
 - Rank the count of landing outcomes between 2010-06-04 & 2017-03-20, in descending order.



https://github.com/KisMyS/Data-Science-Capstone-IBM-Coursera/blob/main/Ex-4-jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

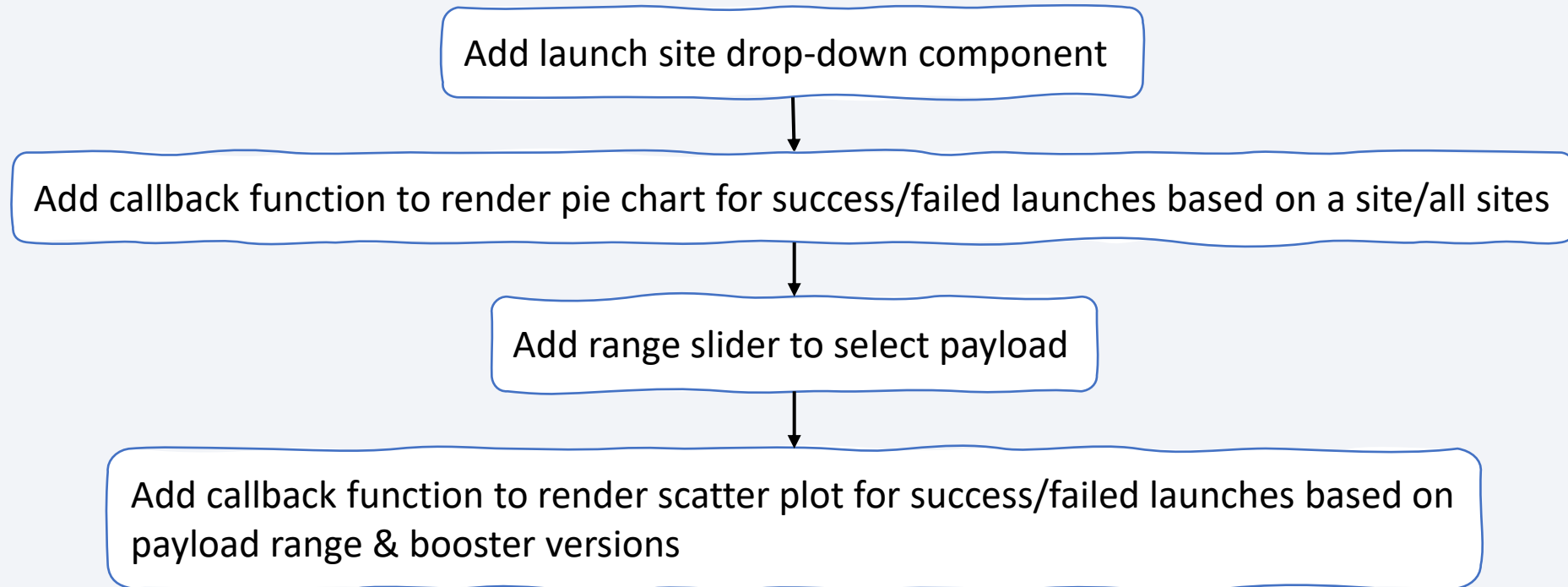
- With interactive map, users can visualize & explore geographical data to better understand the analysis.



<https://github.com/KisMyS/Data-Science-Capstone-IBM-Coursera/blob/main/Ex-6-lab-jupyter-launch-site-location-v2.ipynb>

Build a Dashboard with Plotly Dash

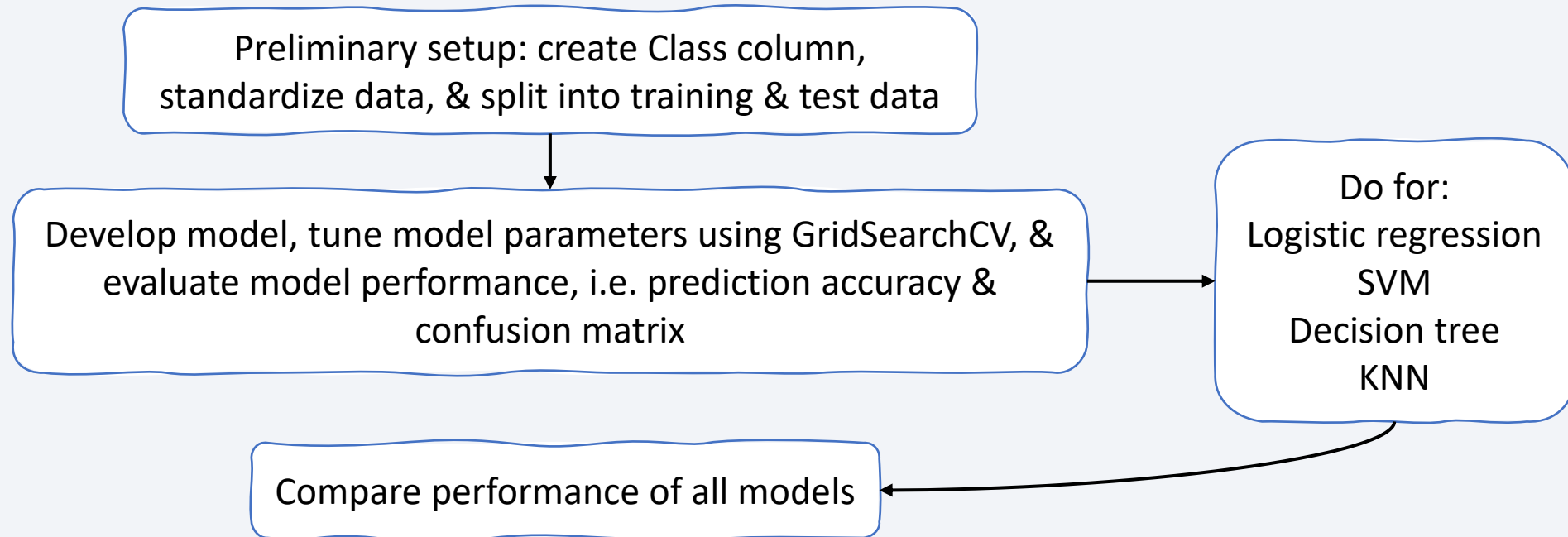
- Users can manipulate & explore summarized data with dashboards.



https://github.com/KisMyS/Data-Science-Capstone-IBM-Coursera/blob/main/Ex-7-spacex_dash_app.py

Predictive Analysis (Classification)

- Lastly, machine learning is used to predict if the first stage lands successfully or not.



https://github.com/KisMyS/Data-Science-Capstone-IBM-Coursera/blob/main/Ex-8-SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

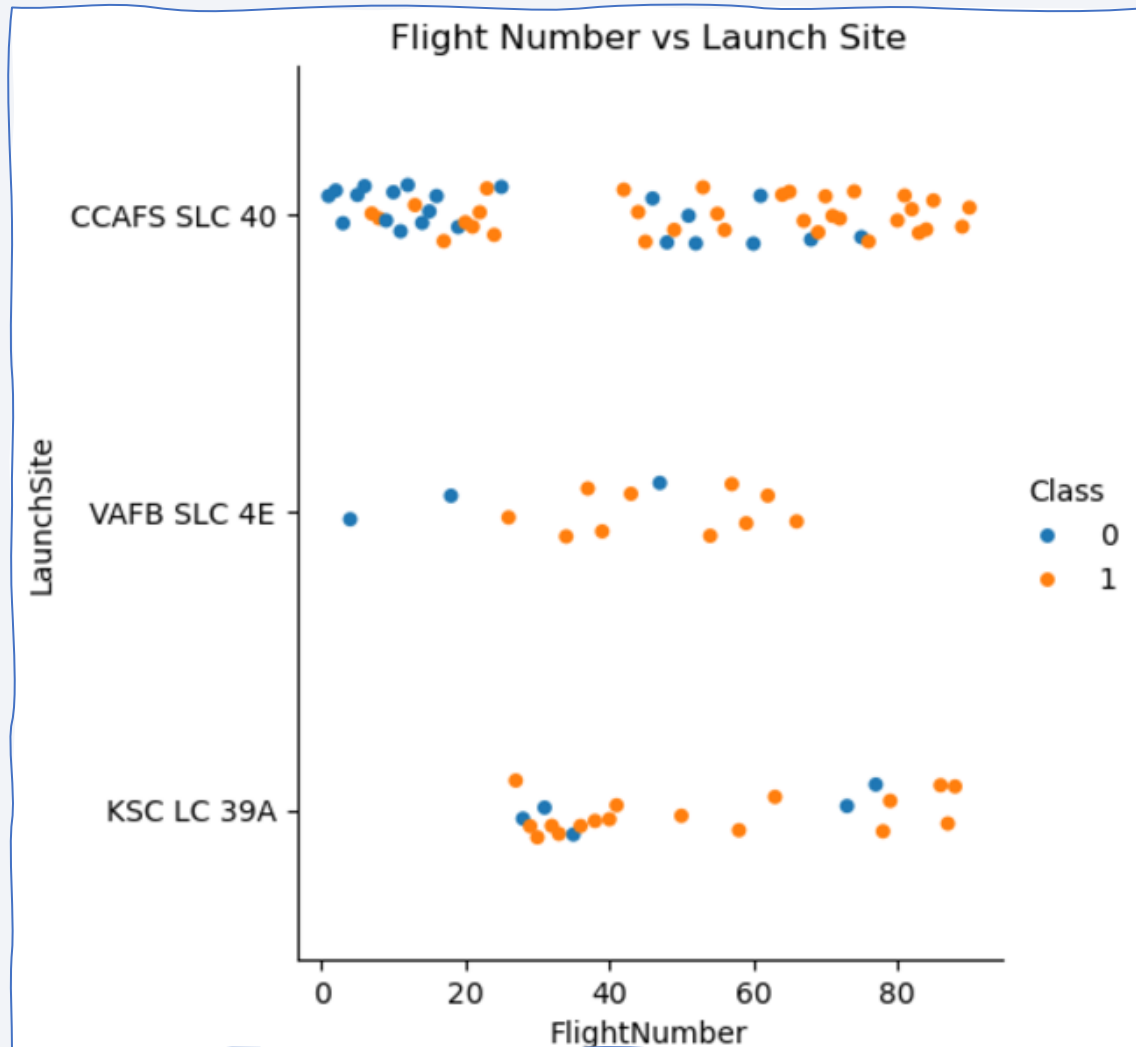
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

The background of the slide is an abstract composition. It features a dark blue base color. Overlaid on this are numerous diagonal streaks in shades of red and cyan. A faint, light blue grid pattern is also visible, particularly in the lower half of the image. The overall effect is dynamic and technological.

Section 2

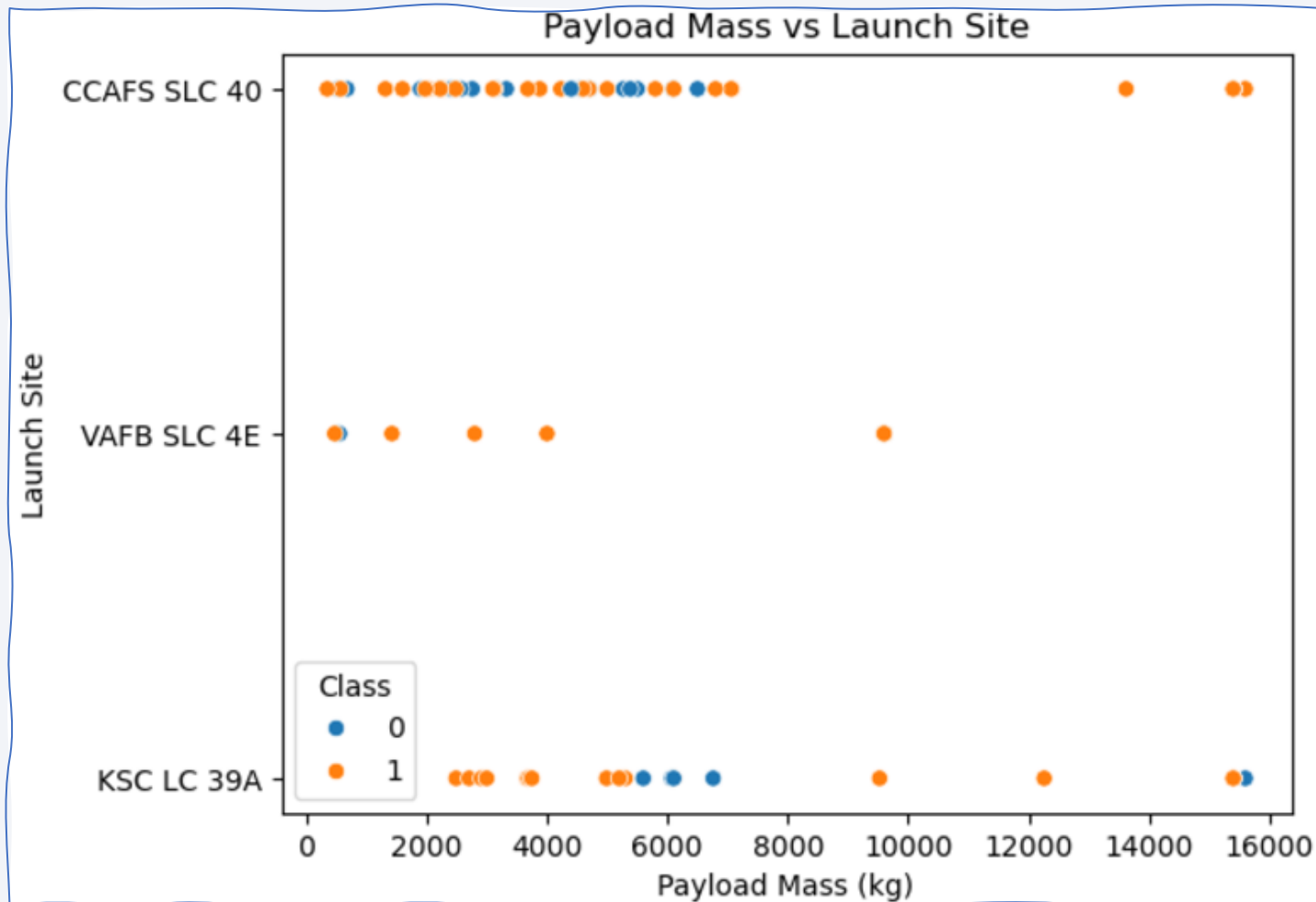
Insights drawn from EDA

Flight Number vs. Launch Site



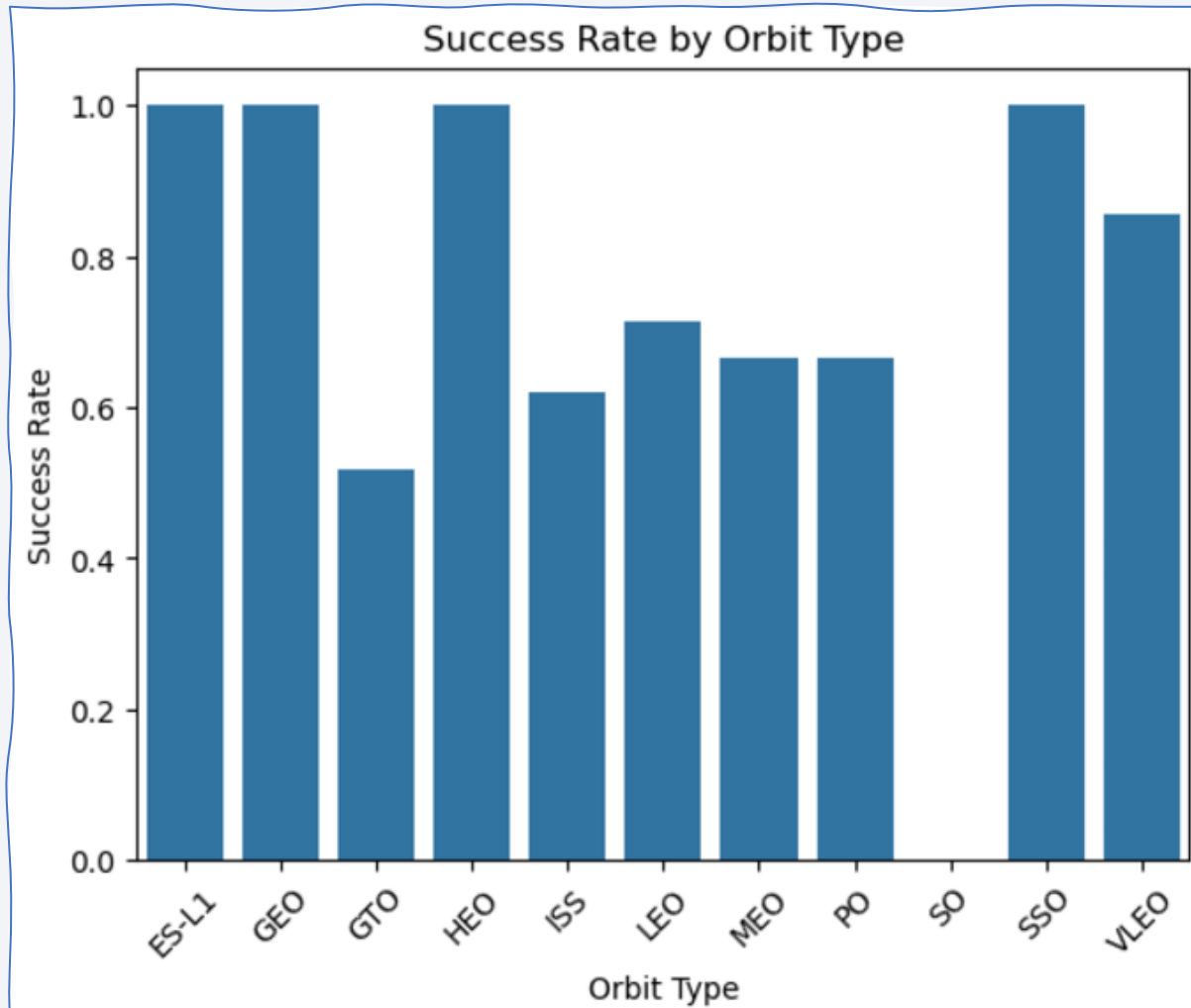
- CCAFS SLC 40 has the most frequent flights among the 3 sites.
- Generally, for each site, the higher the flight number, the higher the success rate.

Payload vs. Launch Site



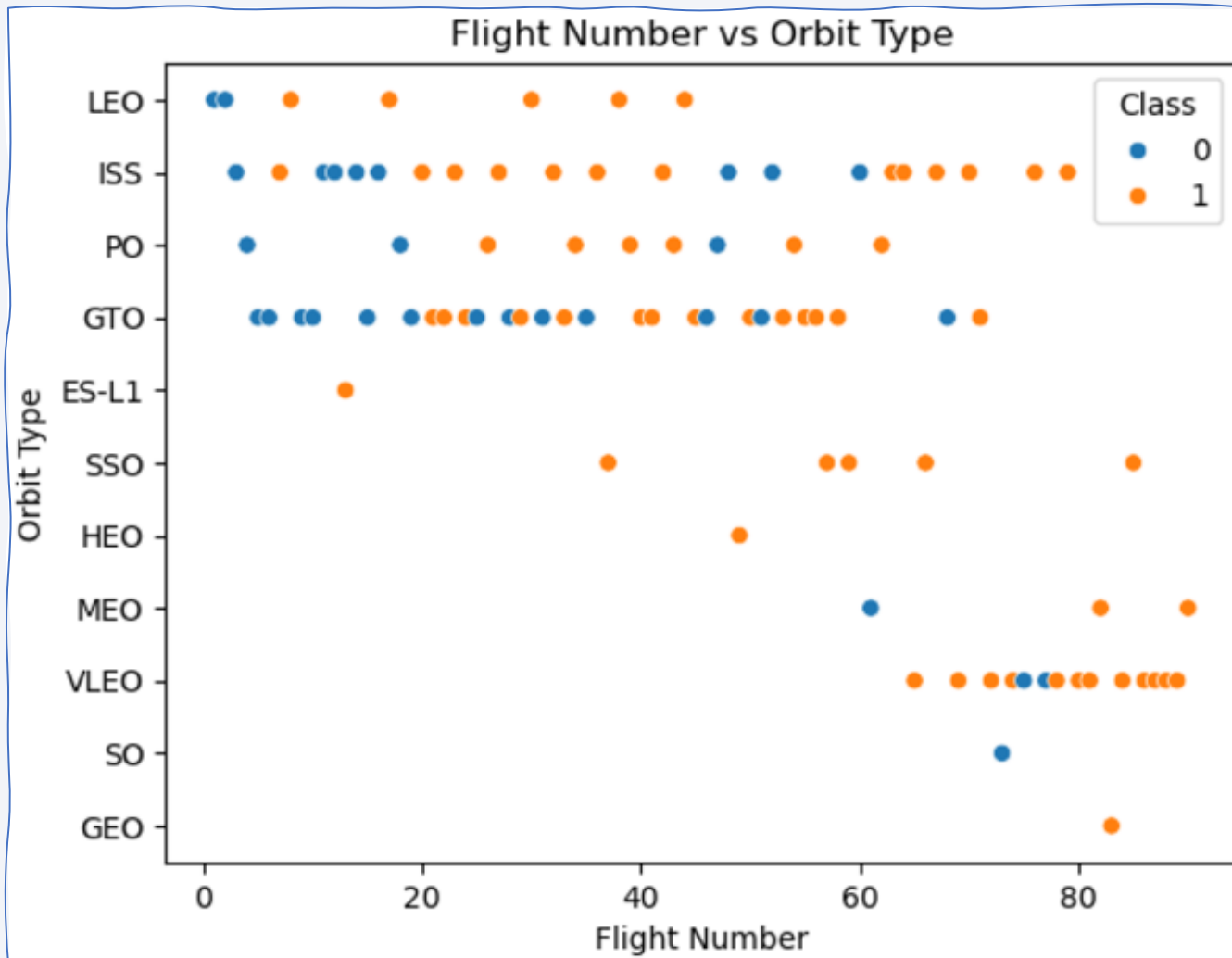
- For all sites, smaller payload mass frequents the launch.
- For VAFB SLC 4E, there are no rockets launched for large payload mass
- There is no clear correlation between success rate & payload mass

Success Rate vs. Orbit Type



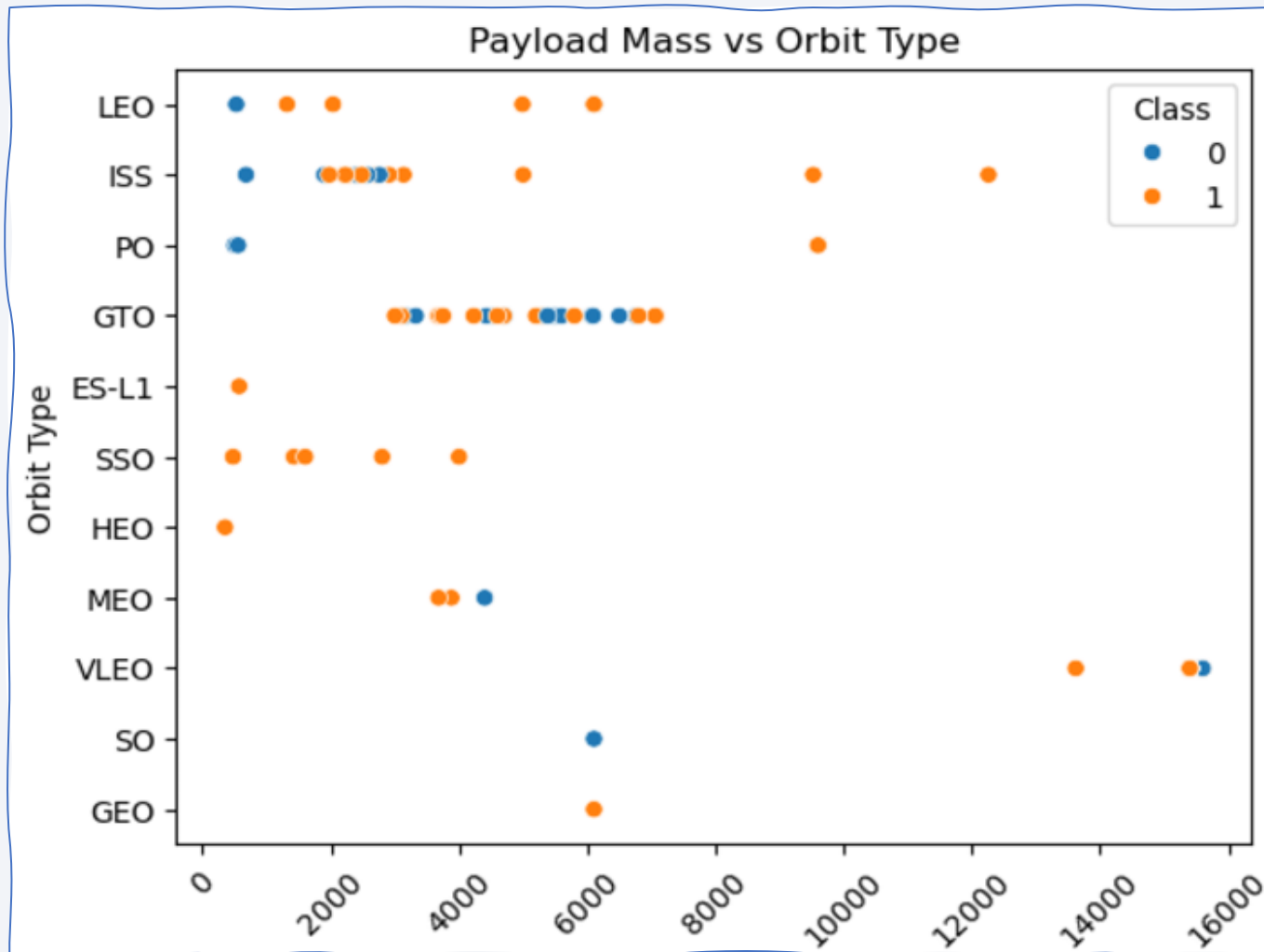
- Orbits with 100% success rate are ES-L1, GEO, HEO & SSO, while SO has 0% success rate
- However, with the exception for SSO, these orbits have only 1 flight recorded
- Other orbit types have the success rate ranging between 50% to about 80%.

Flight Number vs. Orbit Type



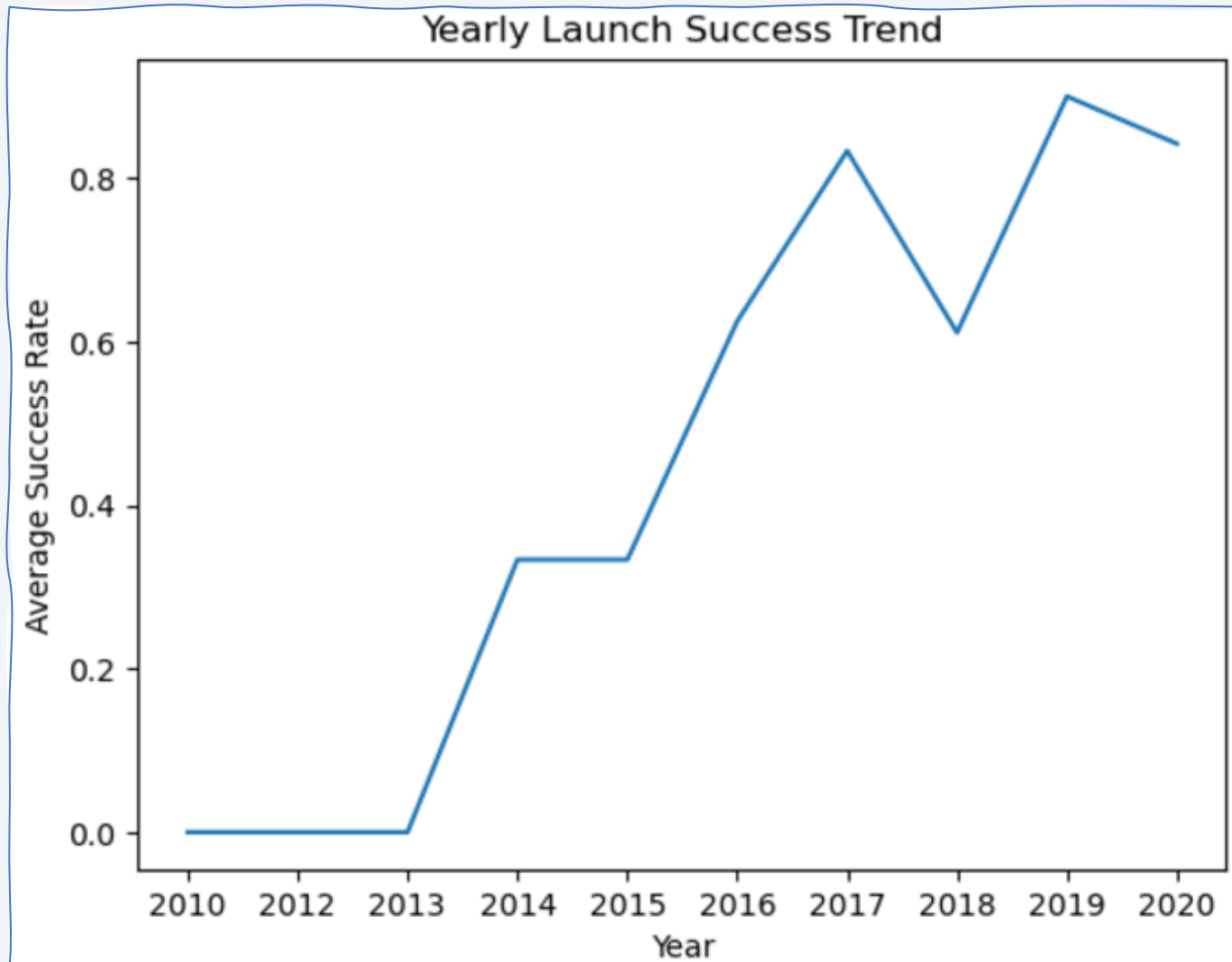
- The 3 most frequent flight are to orbits ISS, GTO & VLEO
- Apart from LEO, PO & MEO, there is no indication that orbit type influences landing success
- There is also no correlation between flight number & success rate with orbit's distance from Earth

Payload vs. Orbit Type



- Heavy payloads have more probability of successful landings for LEO, ISS & PO
- Payload for GTO is in limited range of 2000kg-7000kg
- Payload for ES-L1, SSO & HEO is on the lower range of 0kg to 4000kg, & with 100% success rate
- Not enough information to determine payload influence for MEO, VLEO, SO & GEO

Launch Success Yearly Trend



- Trend shows increase in success rate over the years, with a drop in 2018 & back in trend the next year

All Launch Site Names

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTBL;
```

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

- Use **DISTINCT** to get unique launch site names from **SPACEXTBL**
- Here, we obtained 4 launch sites

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTBL WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	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)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Display 5 records of launch site starting with "CCA"

Total Payload Mass

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE "Customer" =  
'NASA (CRS)';
```

SUM(PAYLOAD_MASS__KG_)
45596

- Calculate the total payload carried by boosters from NASA using SUM()
- The total payload mass is 45,596kg

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG("PAYLOAD_MASS_KG_") FROM SPACEXTBL WHERE  
"Booster_Version" = 'F9 v1.1';
```

AVG(PAYLOAD_MASS_KG_)
2928.4

- Calculate the average payload mass carried by booster version F9 v1.1 using **AVG()**
- The average payload mass is 2,928.4kg

First Successful Ground Landing Date

```
%sql SELECT MIN("Date") FROM SPACEXTBL WHERE "Landing_Outcome" LIKE  
'Success (ground pad)';
```

MIN(Date)
2015-12-22

- Find the dates of the first successful landing outcome on ground pad using **MIN()**, as minimum date intuitively means the first or earliest date
- The date is 22nd of December 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT DISTINCT "Booster_Version" FROM SPACEXTBL WHERE  
"Landing_Outcome" = 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" >  
4000 AND "PAYLOAD_MASS__KG_" < 6000;
```

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

- List the names of boosters which have successfully landed on drone ship & had payload mass between 4000 & 6000 using operators **AND**, **<**, & **>**
- There are 4 booster versions displayed

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT "Landing_Outcome", COUNT(*) AS "Total_Count" FROM  
SPACEXTBL GROUP BY "Landing_Outcome";
```

Landing_Outcome	Total_Count
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	21
No attempt	1
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2

- Calculate the total number of successful & failed mission outcomes using `COUNT()` on the `Landing_Outcome` group
- The landing outcomes & respective total counts are as listed

Boosters Carried Maximum Payload

```
%sql SELECT "Booster_Version" FROM SPACEXTBL WHERE "PAYLOAD_MASS__KG_"  
= (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTBL);
```

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

- List the names of the booster which have carried the maximum payload mass using **MAX()**
- The booster versions are as listed

2015 Launch Records

```
%sql SELECT strftime('%m', "Date") AS "Month", "Landing_Outcome",  
"Booster_Version", "Launch_Site" FROM SPACEXTBL WHERE  
"Landing_Outcome" = 'Failure (drone ship)' AND substr("Date", 1, 4) =  
'2015';
```

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- List the failed landing outcomes in drone ship, their booster versions, & launch site names for 2015 using `strftime()`, `AS`, & `substr()`
- There are 2 of such flights recorded, occurring in January & April 2015

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

```
%sql SELECT "Landing_Outcome", COUNT("Landing_Outcome") AS  
"Outcome_Count" FROM SPACEXTBL WHERE "Date" BETWEEN '2010-06-04' AND  
'2017-03-20' GROUP BY "Landing_Outcome" ORDER BY "Outcome_Count" DESC;
```

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

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between 2010-06-04 & 2017-03-20, in descending order using **COUNT()**, **BETWEEN**, & **ORDER BY ... DESC**
- The landing outcome counts during the period are as listed

A satellite view of Earth from space, showing the curvature of the planet and city lights at night. The background is a deep blue gradient.

Section 3

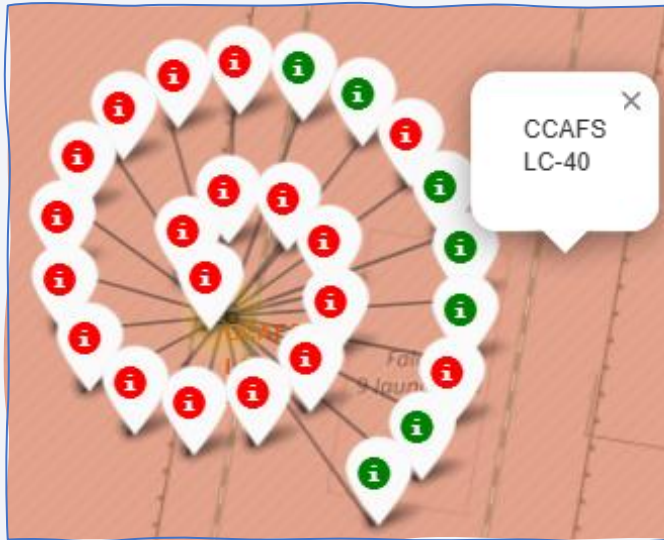
Launch Sites Proximities Analysis

Launch Sites Locations Across the States



- 1 site is in the East Coast & 3 are in the West Coast
- All launch sites are in proximity to the Equator line

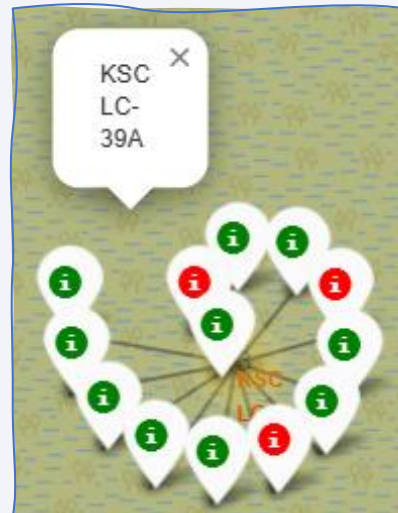
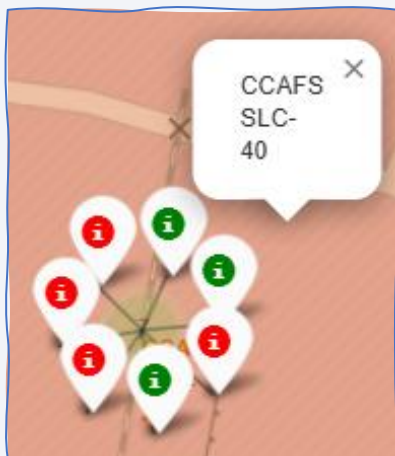
Markers showing success/failed launches



Florida
launch sites



California
launch site



- **Green marker** indicates successful launch, **red marker** indicates failed launch
- CCAFS LC-40 has the most launches
- KSC LC-39A has the highest success rate

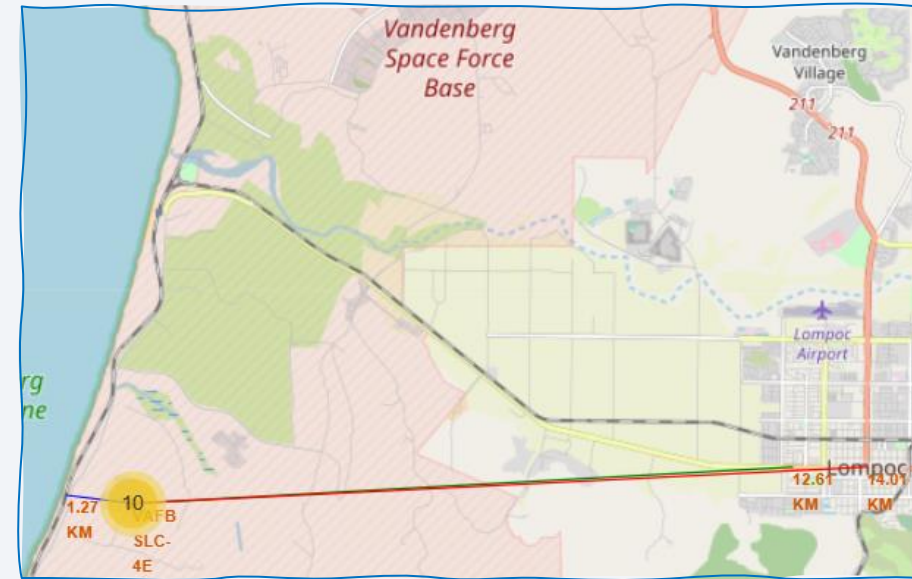
Launch site distance to its proximities

Florida launch sites



Florida East Coast Railway: 21.97 km
Samuel C Phillips Parkway: 0.62 km
Cape Canaveral: 19.26 km

California launch site



Union Pacific Railroad: 1.27 km
CA State Route: 12.61 km
Lompoc: 14.01 km

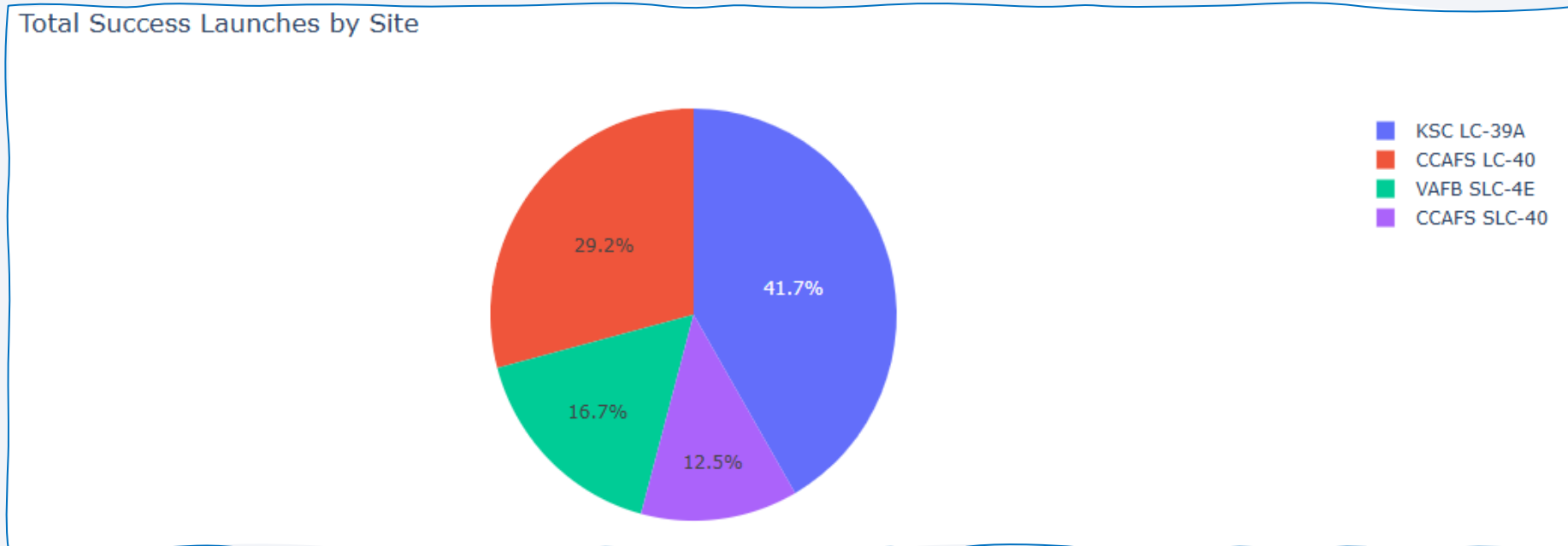
- Launch site may be close to either railway or highway, depending on a site's geographical condition & regional economic activities



Section 4

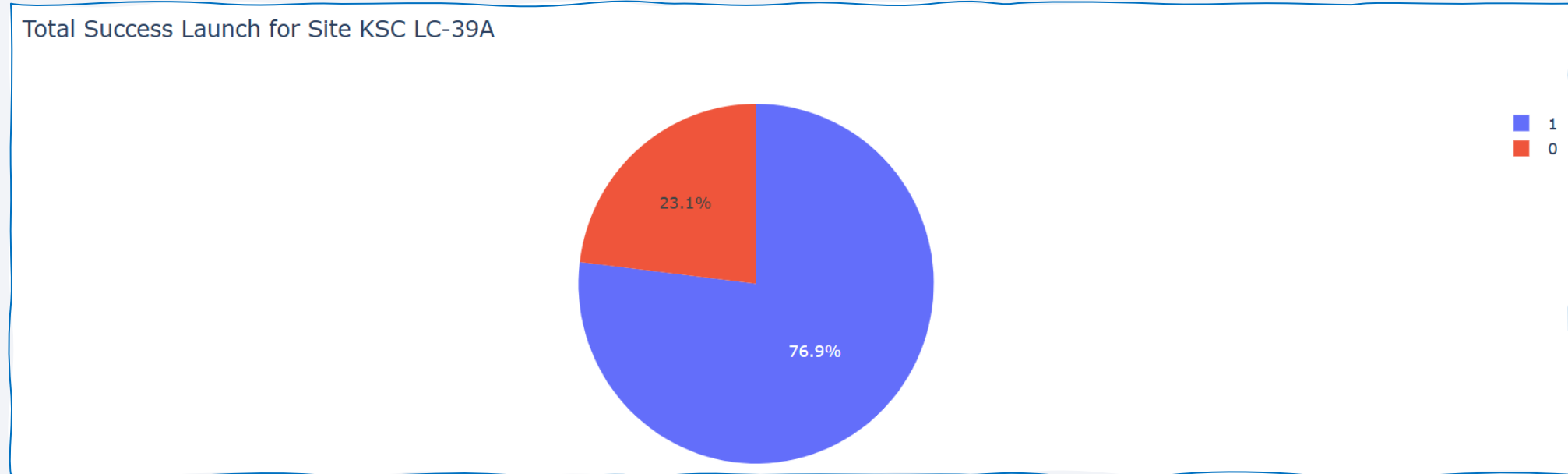
Build a Dashboard with Plotly Dash

Launch Success Count for All Sites



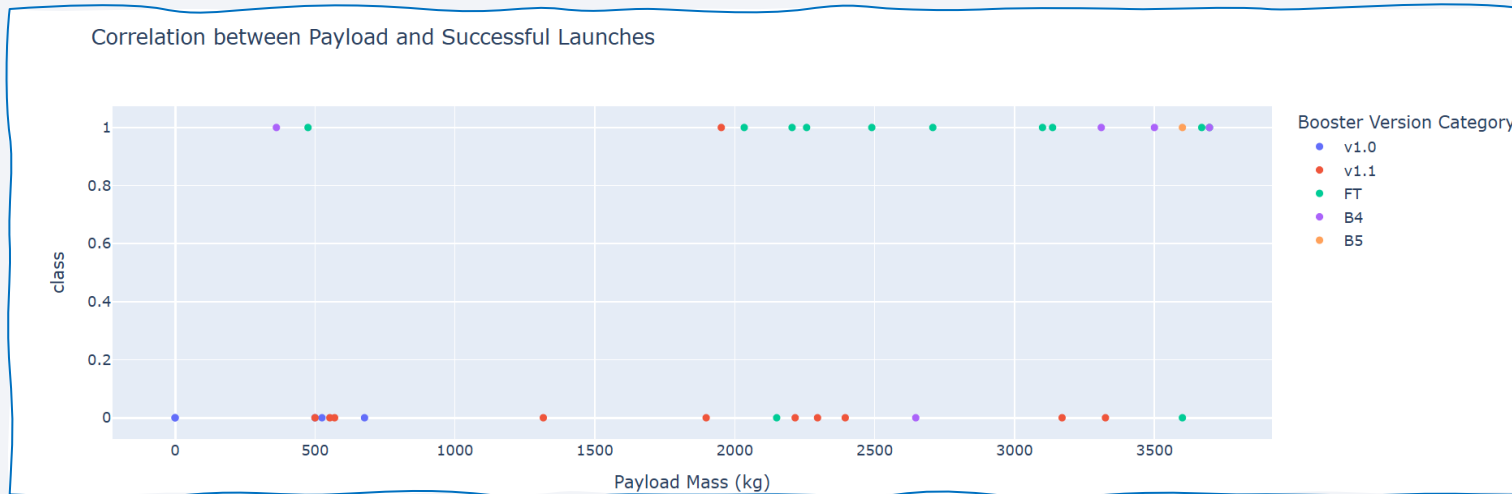
- KSC LC-39A has the highest success count, followed by CCAFS LC-40, VAFB SLC 4E, & CCAFS SLC-40

Highest launch success ratio: KSC LC-39A

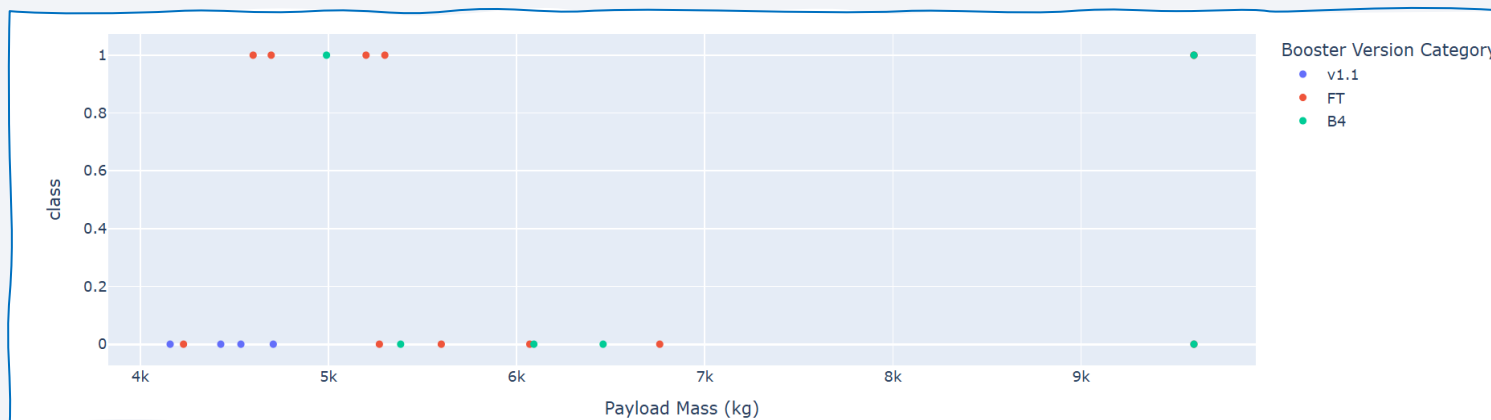


- KSC LC-39A has the success rate of 76.9% & failure rate of 23.1%

Payload vs Launch Outcome for All Sites



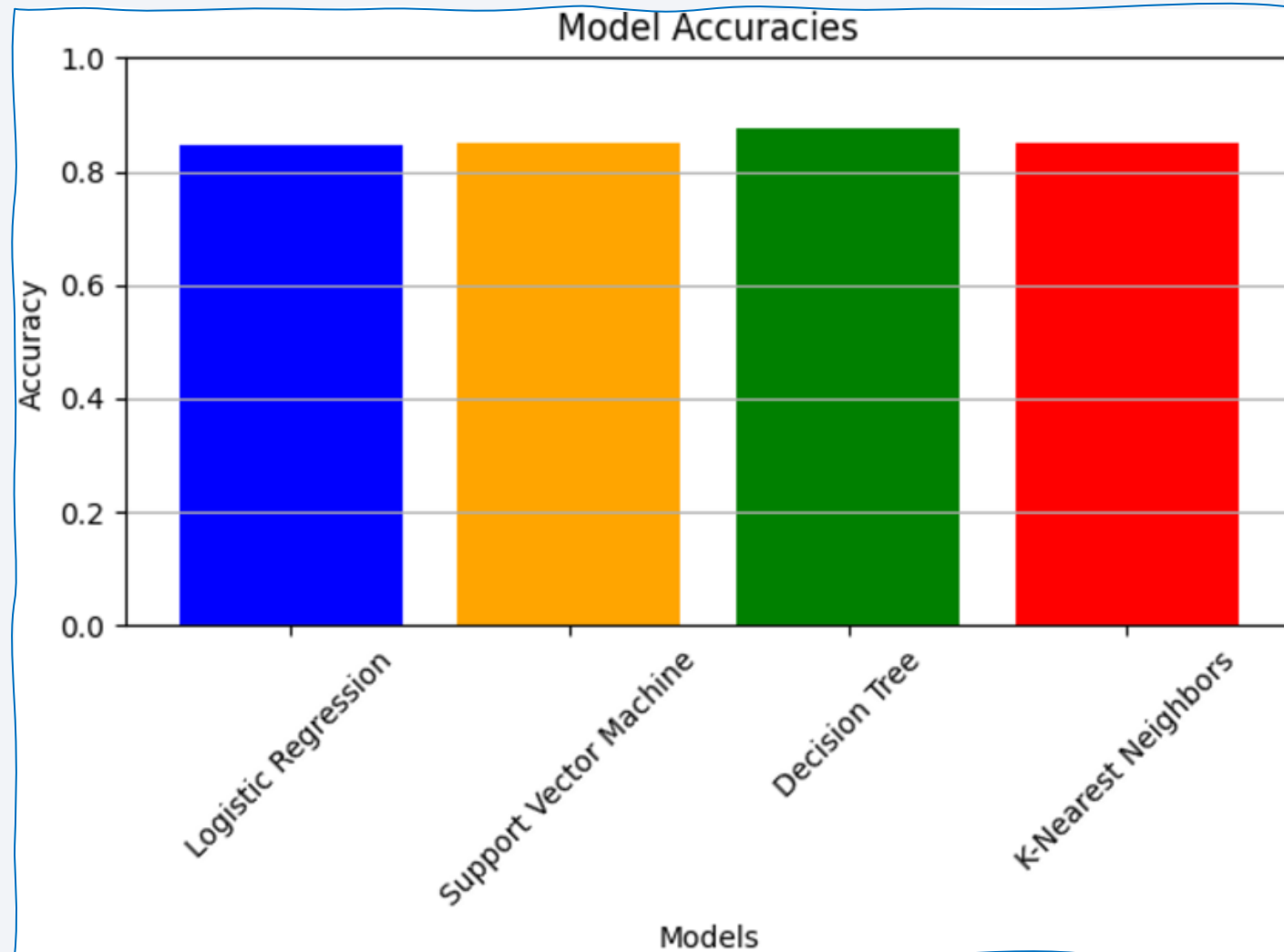
- High success rate sits somewhere for a payload range of 2000kg to 5000kg
- FT booster has the highest success rate with payload range around 2000kg & 3000kg



Section 5

Predictive Analysis (Classification)

Classification Accuracy



Model Accuracies:

Logistic Regression: 0.8464

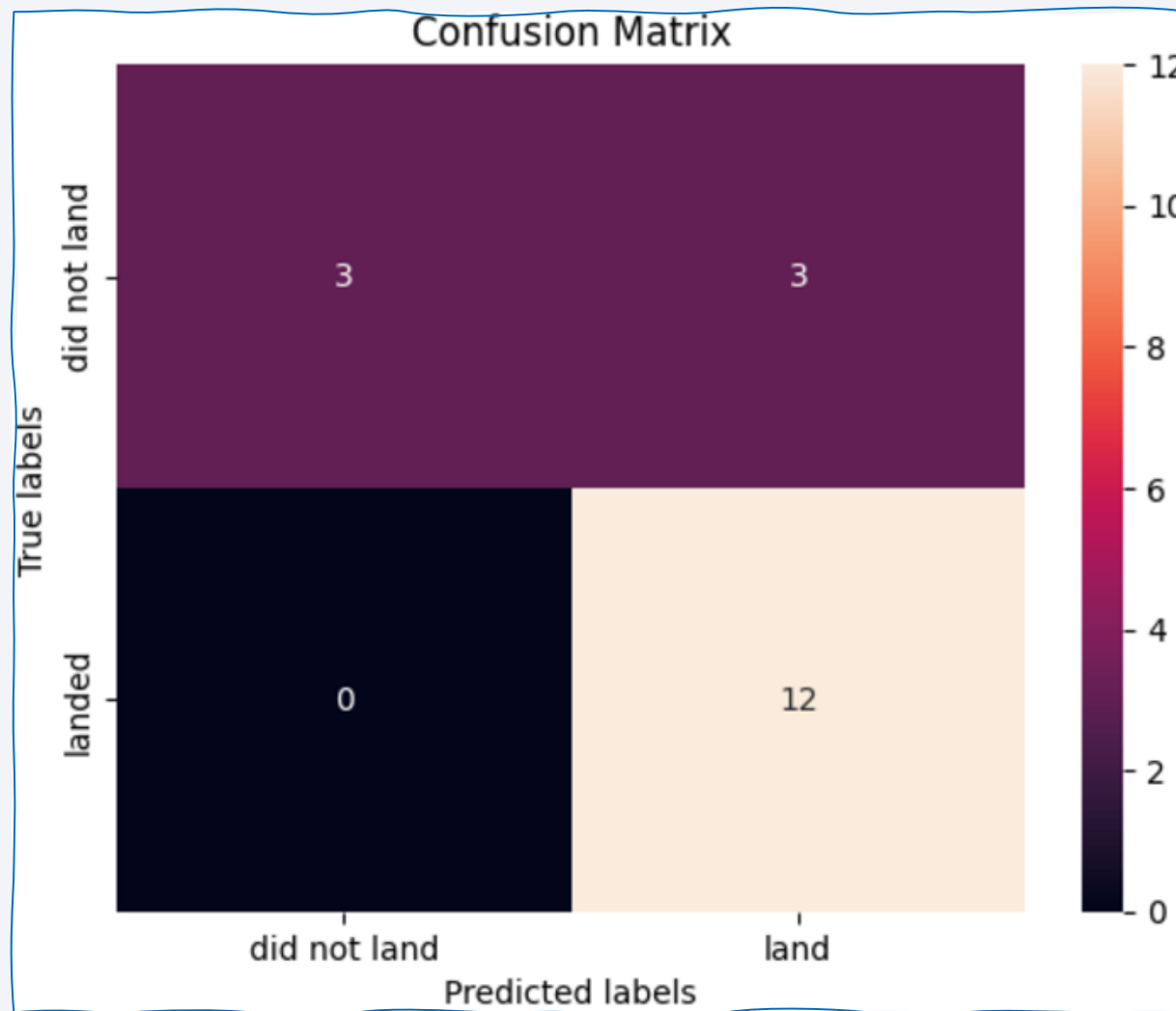
Support Vector Machine: 0.8482

Decision Tree: 0.8750

K-Nearest Neighbors: 0.8482

- Decision tree has the highest accuracy of 0.875

Confusion Matrix



- True positive: 12
 - Model correctly predicted that the rocket landed
- True negative: 3
 - Model correctly predicted that the rocket did not land
- False positive: 3
 - Model predicted that the rocket landed but it did not land
- False negative: 0
 - Model predicted that the rocket did not land but it landed










Conclusions

- Data collection: data is accessed via REST API & webscraping
- Data wrangling: cleaning & appropriately structuring data ensures better analytical experience onwards
- SQL & data visualization helps in understanding the trends in data
- Interactive analysis elevates the exploration experience through maps & dashboard
- Machine learning is used to perform classification with many model selections to choose from, with the aim of achieving appropriately high accuracy

Appendix

- Python code & results are accessible in <https://github.com/KisMyS/Data-Science-Capstone-IBM-Coursera>



 KisMyS Add files via upload	
 Ex-1-jupyter-labs-spacex-data-collection-api-v...	Add files via upload
 Ex-2-jupyter-labs-webscraping.ipynb	Add files via upload
 Ex-3-labs-jupyter-spacex-Data wrangling-v2.ip...	Add files via upload
 Ex-4-jupyter-labs-eda-sql-coursera_sqlite.ipynb	Add files via upload
 Ex-5-jupyter-labs-eda-dataviz-v2.ipynb	Add files via upload
 Ex-6-lab-jupyter-launch-site-location-v2.ipynb	Add files via upload
 Ex-7-spacex_dash_app.py	Add files via upload
 Ex-8-SpaceX_Machine Learning Prediction_Part...	Add files via upload

Thank you!

