IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jens M. W.
2025-02-16

Outline

Executive Summary

Summary of methodologies

- Data Collection and Data Wrangling
- Exploratory Data Analysis (EDA) with SQL and Data Visualization
- Building an Interactive Map with Folium
- Building a Dashboard with Plotly Dash
- Predictive Analysis (Classification)

Summary of all results

- Exploratory Data Analysis results
- Screenshots of the Interactive Map
- Predictive Analysis Results

Introduction

Project background and context

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.

Problems you want to find answers

If we can determine if the first stage will land successfully, we can determine the cost of a launch.

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

Data collection by using SpaceX Rest API as well as using Web Scrapping from Wikipedia

Perform data wrangling

One Hot Encoding data fields for Machine Learning and cleaning of null values and irrelevant columns

Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

Linear Regression (LR), KNN, SVM, DT models have been built and evaluated for the best classifier

Data Collection

Data collection process involved a combination of API requests from SpaceX REST API to get launch data and Web Scraping data from a table in SpaceX’s Wikipedia entry.

Sources:

SpaceX API (structured launch data)
Wikipedia (historical launch records via web scraping)

Process:

API: Extract, normalize, clean, and preprocess data
Web Scraping: Parse HTML, extract tables, clean & structure data

Output:

Processed datasets stored as CSV and HTML tables for further analysis.

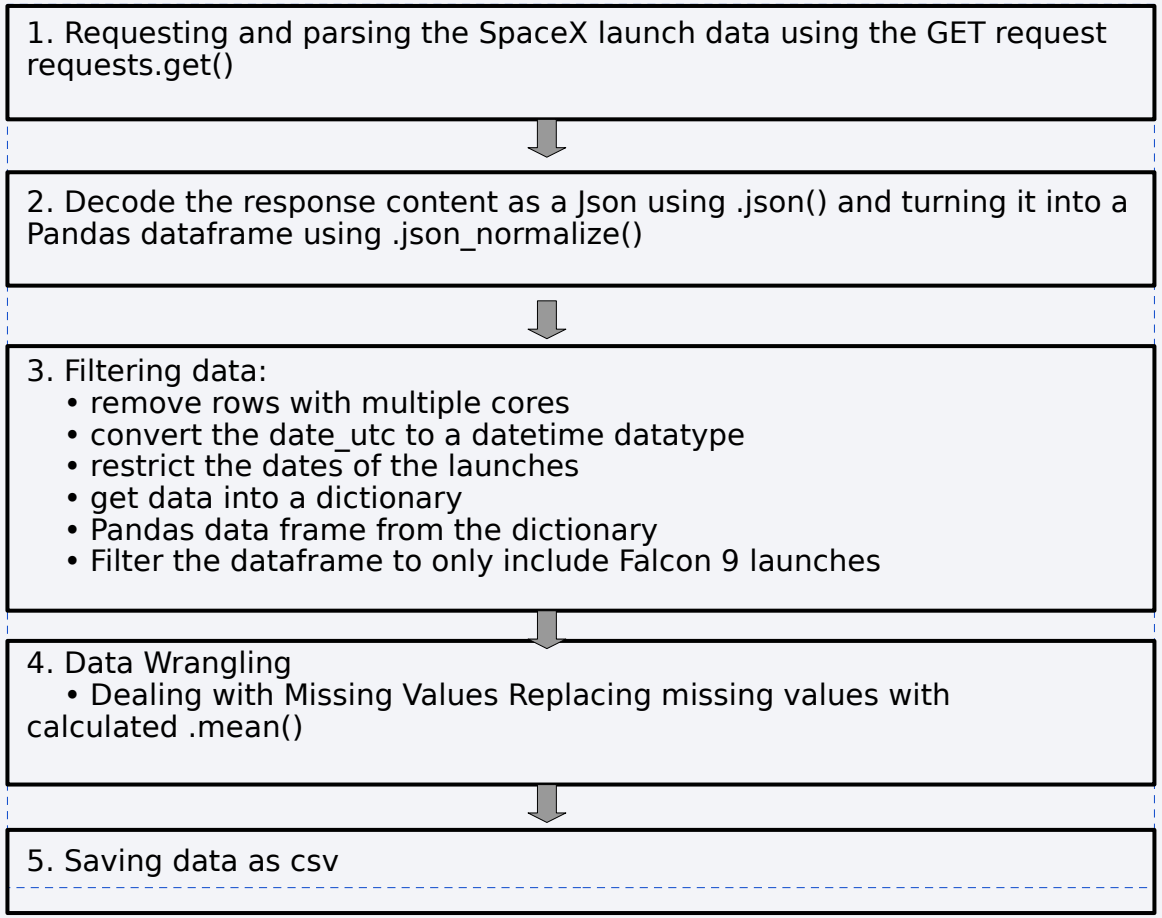
Data Collection – SpaceX API

- Launch data is gathered using the SpaceX REST API

  Data Columns used:

  Data - BoosterVersion - PayloadMass – Orbit – LaunchSite - Outcome - Flights - GridFins – Reused - Legs - LandingPad - Block - ReusedCount - Serial - Longitude - Latitude

- https://github.com/jmw-geo/capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

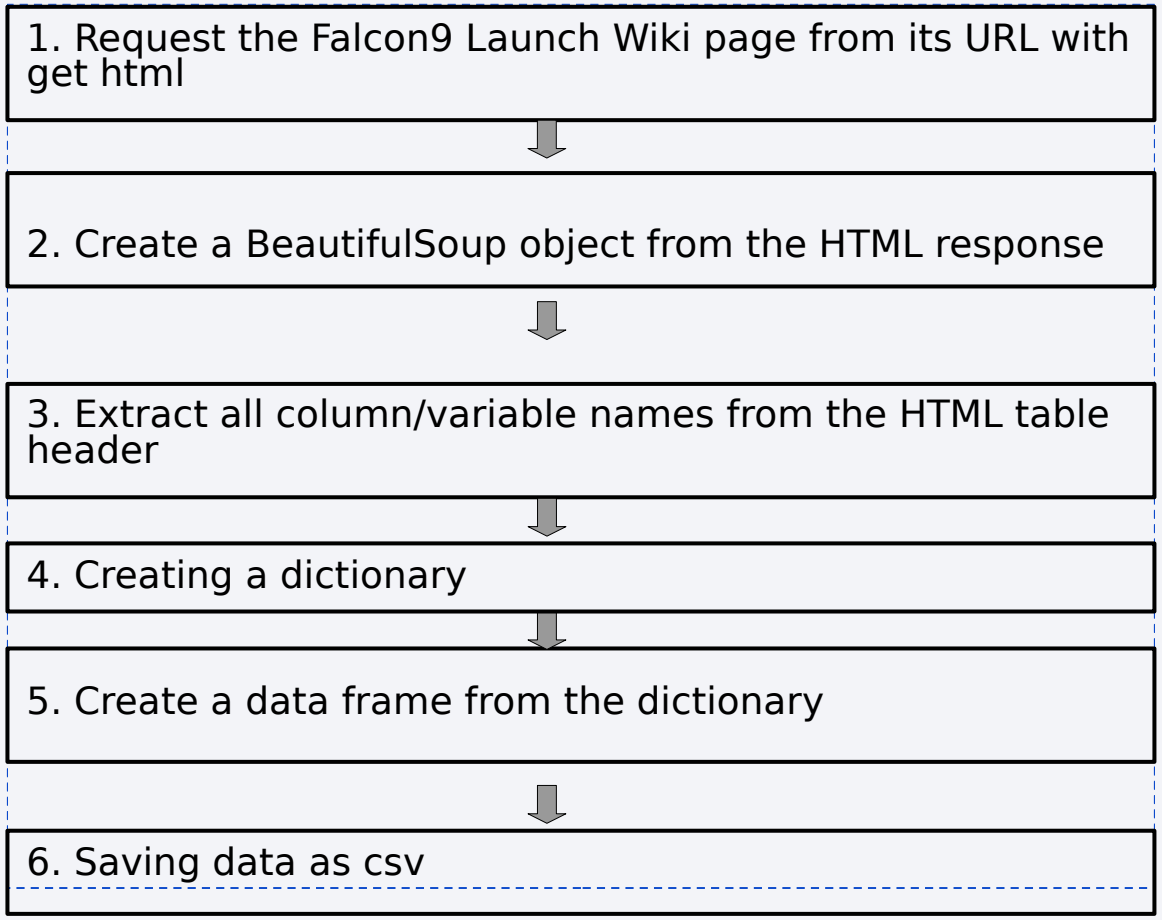


Data Collection - Scraping

- Web scraping Falcon9 historical Launch records from Wiki page

  Data Columns are obtained by using Wikipedia Web Scraping:
  Flight No. - Launch site – Payload – PayloadMass – Orbit – Customer – Launch outcome - Version Booster - Booster landing – Date - Time

- https://github.com/jmw-geo/capstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

Objectives:

Perform exploratory Data Analysis and determine Training Labels:

- Exploratory Data Analysis
- Determine Training Labels

Methods:

- Identify and calculate the percentage of the missing values in each attribute
- Calculation the number of launches on each site
- Calculation of the number and occurrence of each orbit
- Calculation of the number and occurrence of mission outcome of the orbits
- Creation of a landing outcome label from the Outcome column
- Exporting data to csv

https://github.com/jmw-geo/capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Summarize what charts were plotted and why you used those charts:

- relationship between Flight Number and Payload Mass

- relationship between Flight Number and Launch Site

- relationship between Payload Mass and Launch Site

- relationship between success rate of each orbit type

- relationship between Flight Number and Orbit type

- relationship between Payload Mass and Orbit type

- launch success yearly trend

Scatter plots show the relationship between variables. If a relationship exists, they could be used in machine learning model.
Bar charts show comparisons among discrete categories to show the relationship between the specific categories being compared and a measured value.
Line charts show trends in data over time (time series).

https://github.com/jmw-geo/capstone/blob/main/edadataviz.ipynb

EDA with SQL

Performed SQL queries:

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string ‘CCA'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date when the first successful landing outcome in ground pad was achieved
- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster versions which have carried the maximum payload mass
- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015
- Ranking 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

https://github.com/jmw-geo/capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Markers all launch sites on a map to get the following information:

- Are all launch sites in proximity to the Equator line?
- Are all launch sites in very close proximity to the coast?

Color-coded markers of the success (green) / failed (red) launches for each site on the map to be able to easily identify which launch sites have relatively high success rates.

Calculation of the distances between a launch site to its proximities (in this case: Vandenberg California VAFB SLC-4E launch site) like Railway, Highway, Coastline and closest City:

- Are all sites placed at a safe distance away from cities ?
- Are all launch sites close to the coast ?
- Are all launch sites close to railway

https://github.com/jmw-geo/capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Launch Sites Drop-down List:

- Added a drop-down list to enable Launch Site selection.

Pie Chart showing Success Launches (All Sites/Certain Site):

- Added a pie chart to show the total successful launches count for all sites and the Success vs. Failed counts for the site, if a specific Launch Site was selected.

Slider of Payload Mass Range:

- Added a slider to select Payload range.

Scatter Chart of Payload Mass vs. Success Rate for the different Booster Versions:

- Added a scatter chart to show the correlation between Payload and Launch Success.

https://github.com/jmw-geo/capstone/blob/main/spacex_dash_app.py

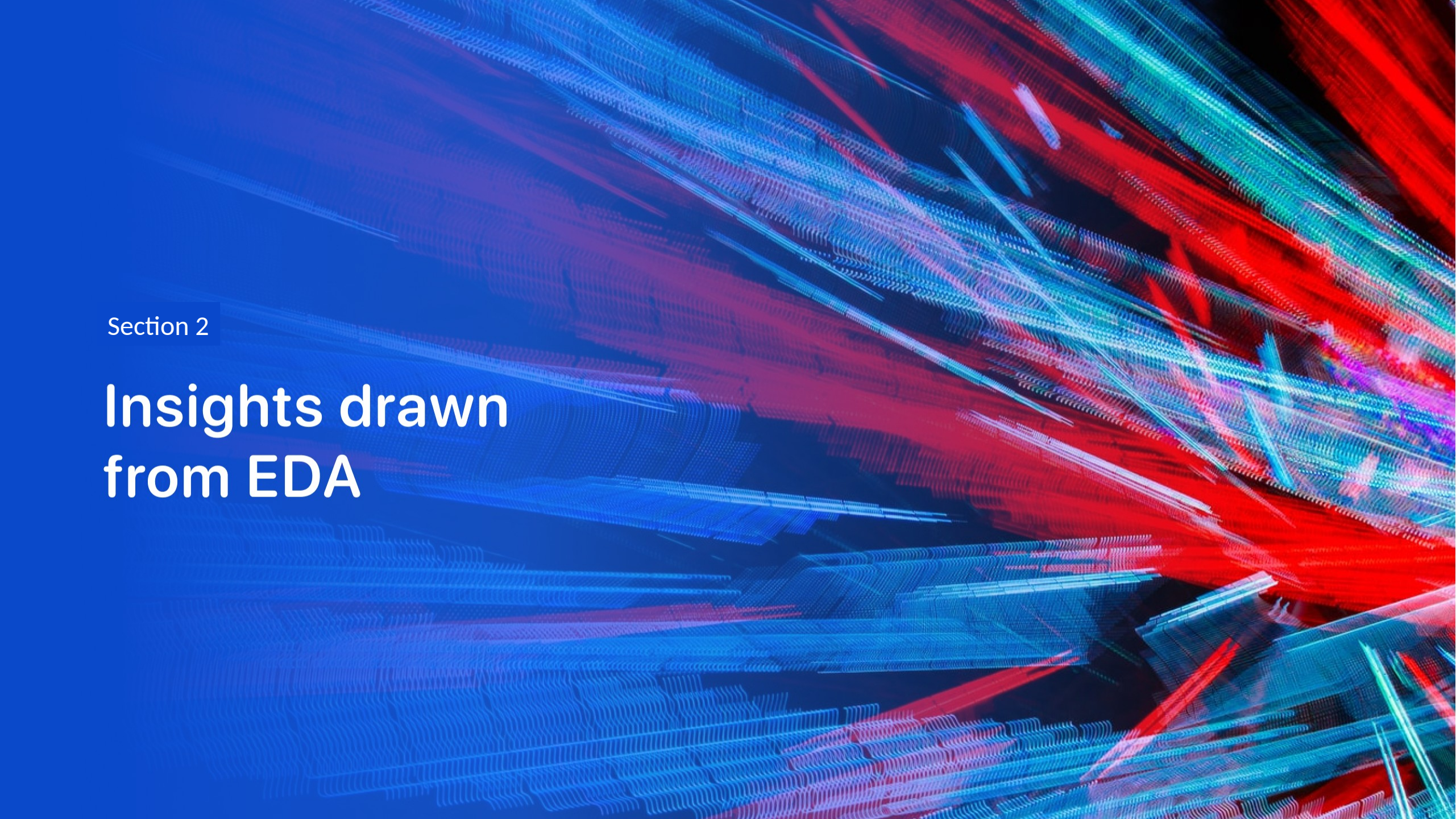
Predictive Analysis (Classification)

1. Create a NumPy array from the column Class in data
2. Standardize the data in X
3. Use the function train_test_split to split the data X and Y into training and test data
4. Create a GridSearchCV object logreg_cv with cv = 10 and fit the object to find the best parameters from the dictionary parameters.
5. Applying GridSearchCV on LogReg, SVM, Decision Tree, and KNN models
6. Calculating the accuracy on the test data using the method .score() for all models
7. Examining the confusion matrix for all models
8. Finding the method that performs best

https://github.com/jmw-geo/capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

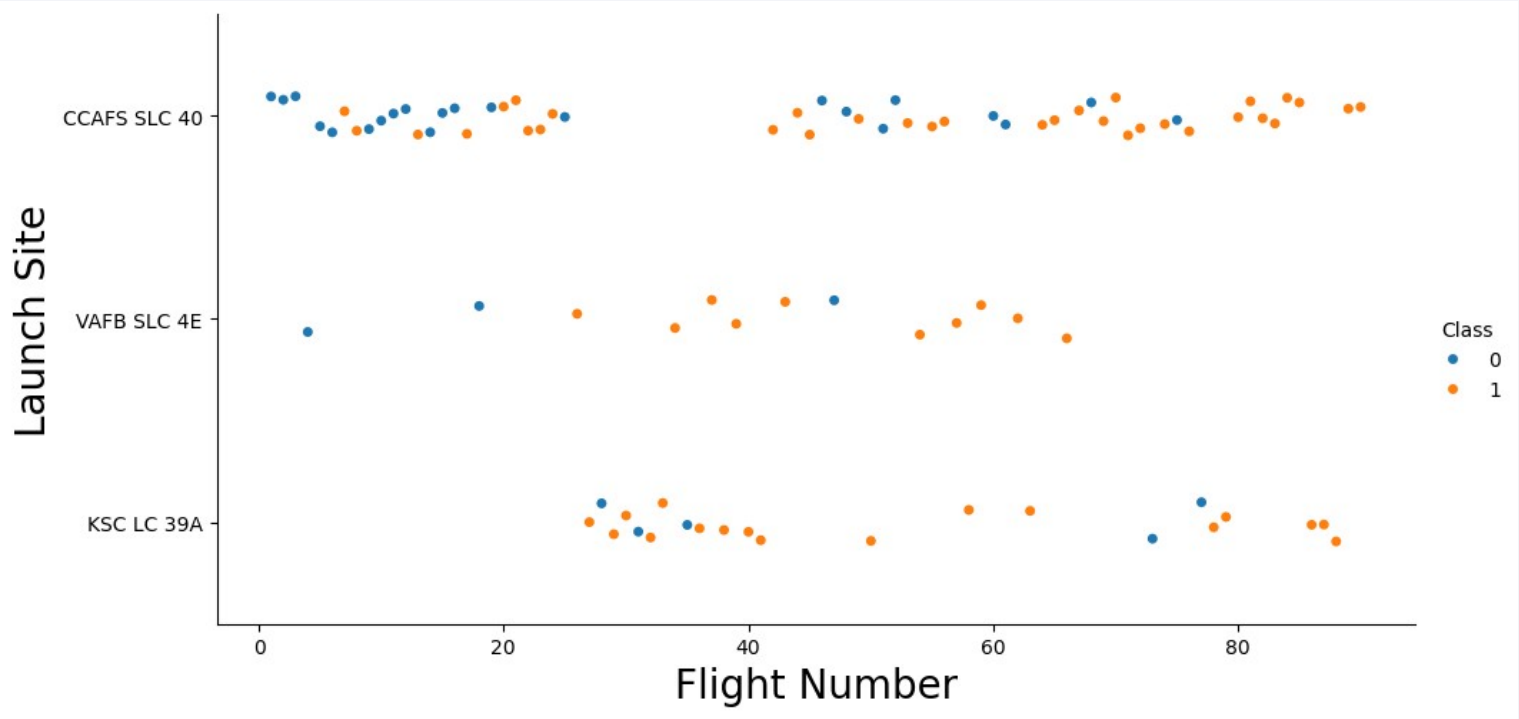
Results

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


Section 2
Insights drawn from EDA

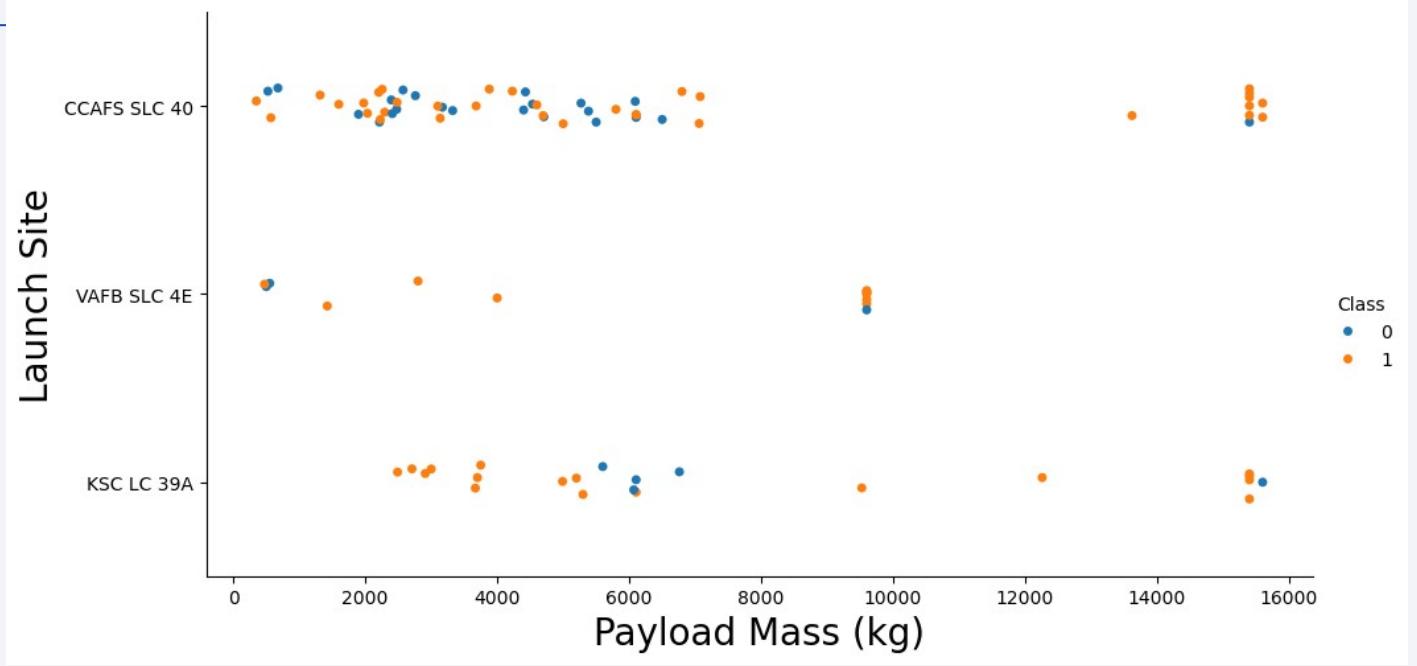
Flight Number vs. Launch Site

- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- The earliest flights all failed, the latest flights almost all succeeded.
- It can be assumed that a new launch has a higher rate of success.



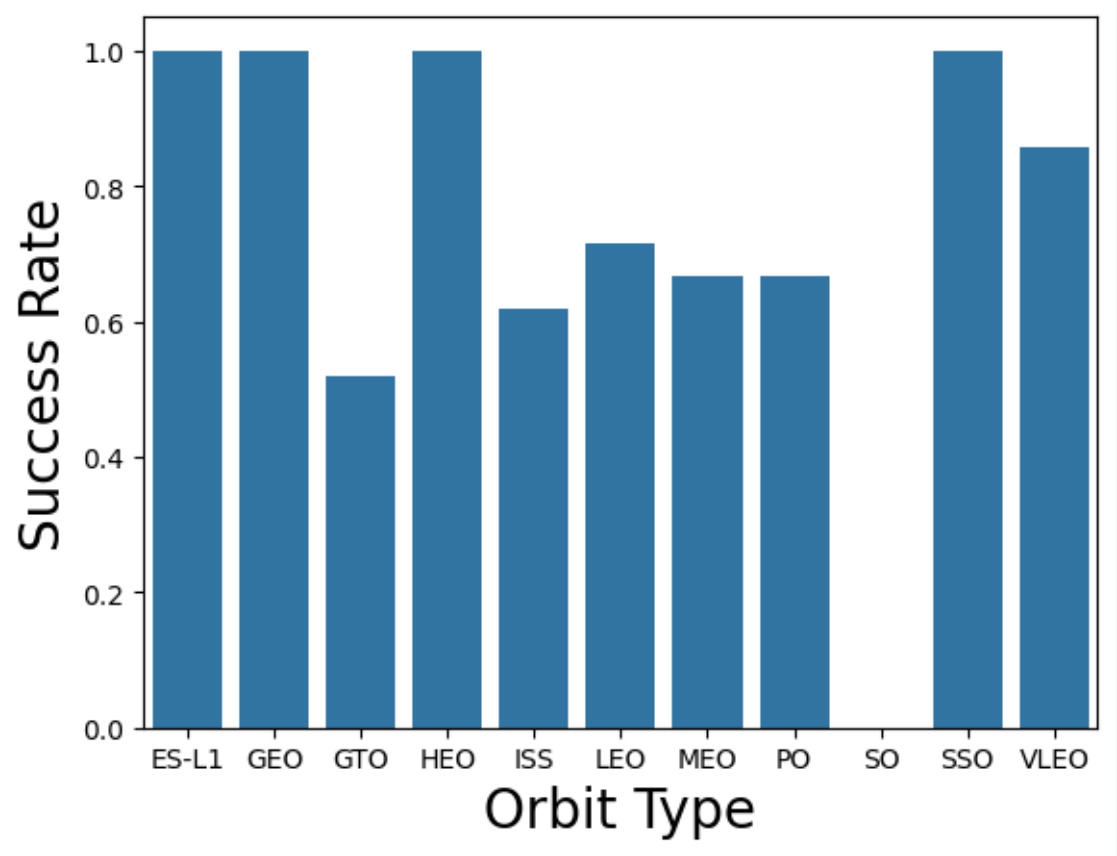
Payload vs. Launch Site

- Most of the launches with payload mass over 8000 kg were successful.
- Higher payloads (8,000+ kg) appear mostly at CCSFSSLC40 & KSCLC39A and tend to have more successful landings.
- VAFBSLC4E is the least used site.



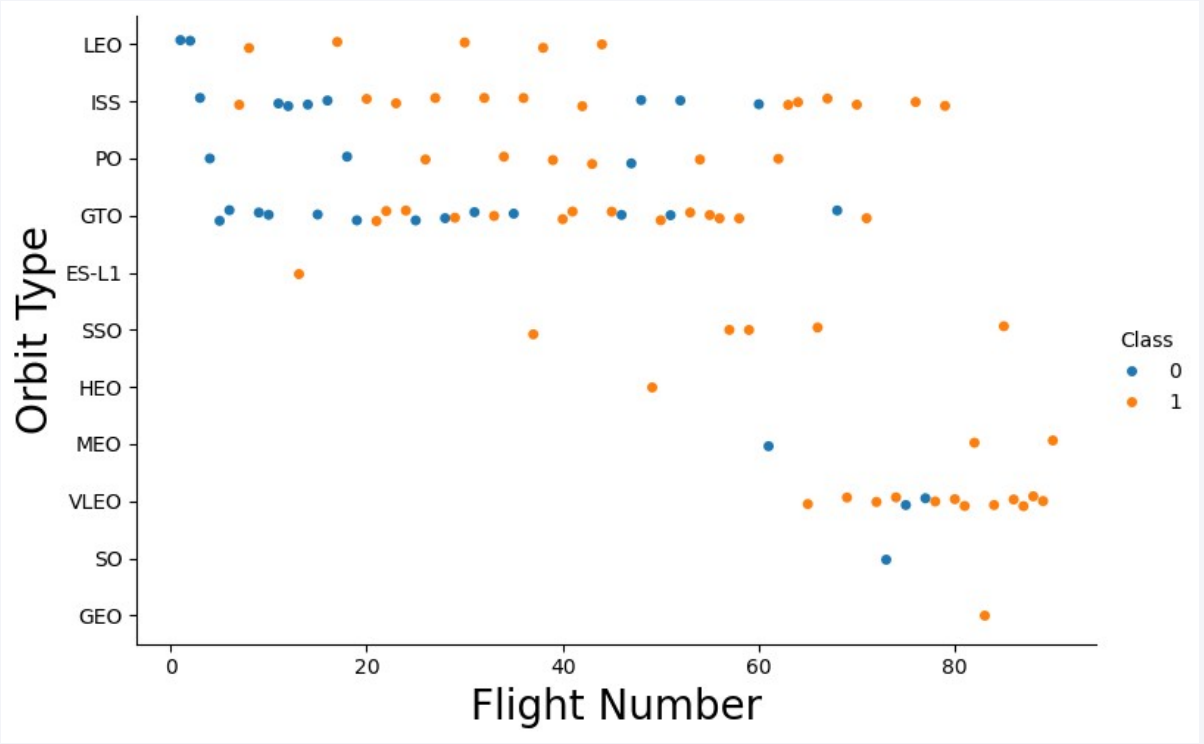
Success Rate vs. Orbit Type

- The orbit types of ES-L1, GEO, HEO and SSO show the highest success rate with 100%
- SO has a 0% success rate
- The other orbits show a success rate between 50% and 85%



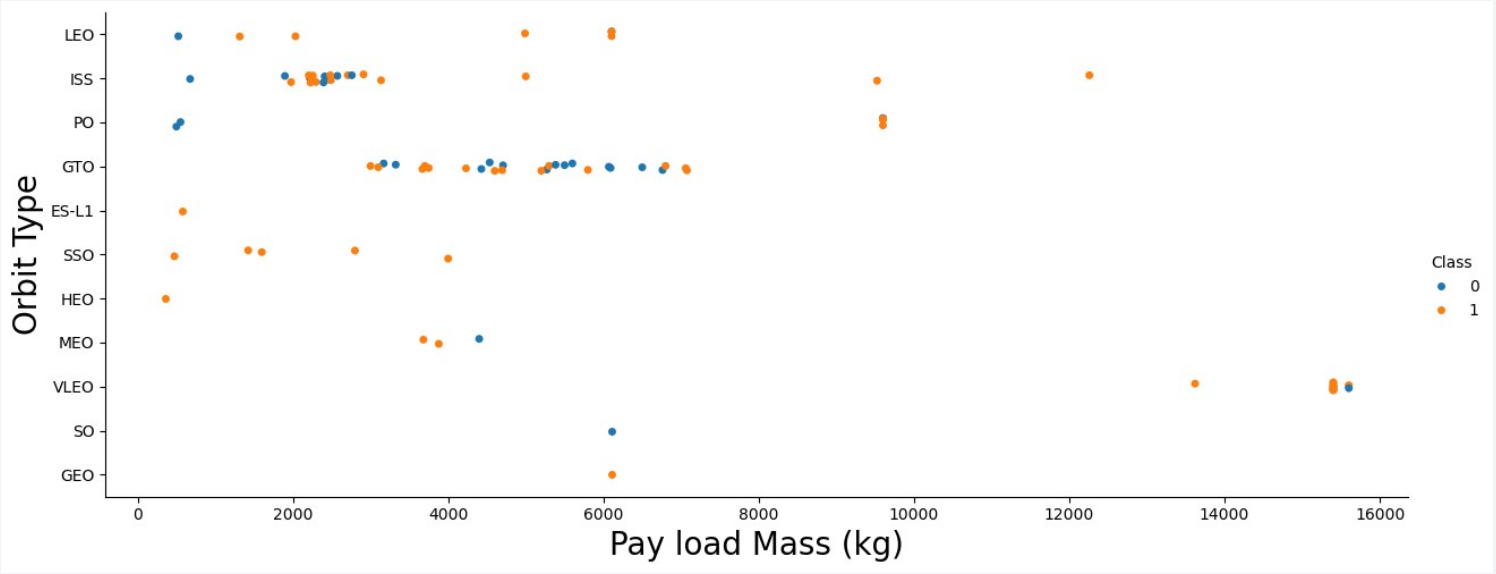
Flight Number vs. Orbit Type

- With more launches per site, the success rate increased significantly
- Orbit VLEO shows the highest success rate
- There seem to be be no clear relation between flight numbers and success rates at orbit GTO



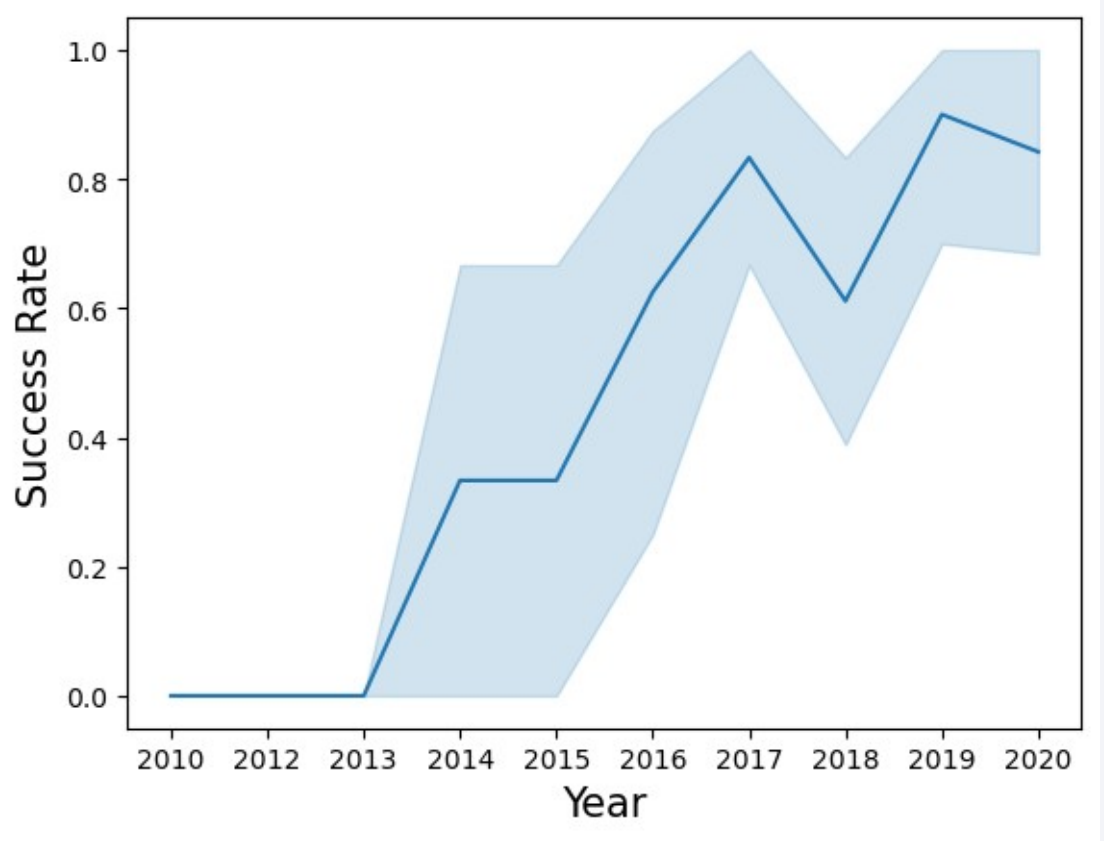
Payload vs. Orbit Type

- ISS & LEO: Higher payloads tend to succeed more often, while smaller payloads experience more failures.
- GTO shows mixed success rate.
- Over all, there seem to be no clear correlation between payload mass and success rate



Launch Success Yearly Trend

- The success rate kept increasing since 2013 till 2020 with a temporary decline in 2018



All Launch Site Names

[34]:
```
#dataframe = pd.read_sql_query("select * from SPACEXTABLE;", con)

#print the dataframe
#dataframe
dataframe = pd.read_sql('select distinct Launch_Site from SPACEXTABLE;', con)
dataframe
#dataframe = pd.read_sql('select distinct Landing_Outcome from SPACEXTABLE;', con)
#dataframe
```

[34]:

	Launch_Site
0	CCAFS LC-40
1	VAFB SLC-4E
2	KSC LC-39A
3	CCAFS SLC-40

Displaying the names of the unique launch sites in the space mission.

Launch Site Names Begin with 'CCA'

Displaying the site names beginning with ‚CCA‘ (first 5 records)

[14]:
```
dataframe = pd.read_sql("select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5", con)
dataframe
```

[14]:

	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
0	2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	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
3	2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

```
[15]: dataframe = pd.read_sql("select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer='NASA (CRS)'", con)
      dataframe
```

[15]:

	sum(PAYLOAD_MASS__KG_)
0	45596

Displaying the total payload mass carried by boosters launched by NASA (CRS)

Average Payload Mass by F9 v1.1

```
[16]: dataframe = pd.read_sql("select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version='F9 v1.1'", con)
      dataframe
```

[16]:

	avg(PAYLOAD_MASS__KG_)
0	2928.4

Average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

[17]:
```
dataframe = pd.read_sql("select min(Date) from SPACEXTABLE where Landing_Outcome='Success (ground pad)'", con)
dataframe
```

[17]:

	min(Date)
0	2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

[18]:
```
dataframe = pd.read_sql("select distinct Booster_Version from SPACEXTABLE where Landing_Outcome='Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 an
dataframe
```

[18]:

	Booster_Version
0	F9 FT B1022
1	F9 FT B1026
2	F9 FT B1021.2
3	F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

[21]:
```
dataframe = pd.read_sql("select substr(Mission_Outcome,1,7) as Mission_Outcome, count(*) from SPACEXTABLE group by 1", con)
dataframe
```

[21]:

	Mission_Outcome	count(*)
0	Failure	1
1	Success	100

Boosters Carried Maximum Payload

```
[23]: dataframe = pd.read_sql("select distinct Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)", con)
      dataframe
```

[23]:

	Booster_Version
0	F9 B5 B1048.4
1	F9 B5 B1049.4
2	F9 B5 B1051.3
3	F9 B5 B1056.4
4	F9 B5 B1048.5
5	F9 B5 B1051.4
6	F9 B5 B1049.5
7	F9 B5 B1060.2
8	F9 B5 B1058.3
9	F9 B5 B1051.6
10	F9 B5 B1060.3
11	F9 B5 B1049.7

2015 Launch Records

```
[32]: dataframe = pd.read_sql("select distinct substr(Date, 6,2), Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE where Landing_Outcome ='Failure (
      dataframe
```

[32]:

	substr(Date, 6,2)	Landing_Outcome	Booster_Version	Launch_Site
0	01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
1	04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Listing the failed landing outcomes in 2015

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

[33]:
```
dataframe = pd.read_sql("select Landing_Outcome, count(*) from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order b
dataframe
```

[33]:

	Landing_Outcome	count(*)
0	No attempt	10
1	Success (drone ship)	5
2	Failure (drone ship)	5
3	Success (ground pad)	3
4	Controlled (ocean)	3
5	Uncontrolled (ocean)	2
6	Failure (parachute)	2
7	Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Global Map of SpaceX Launch Sites in the USA

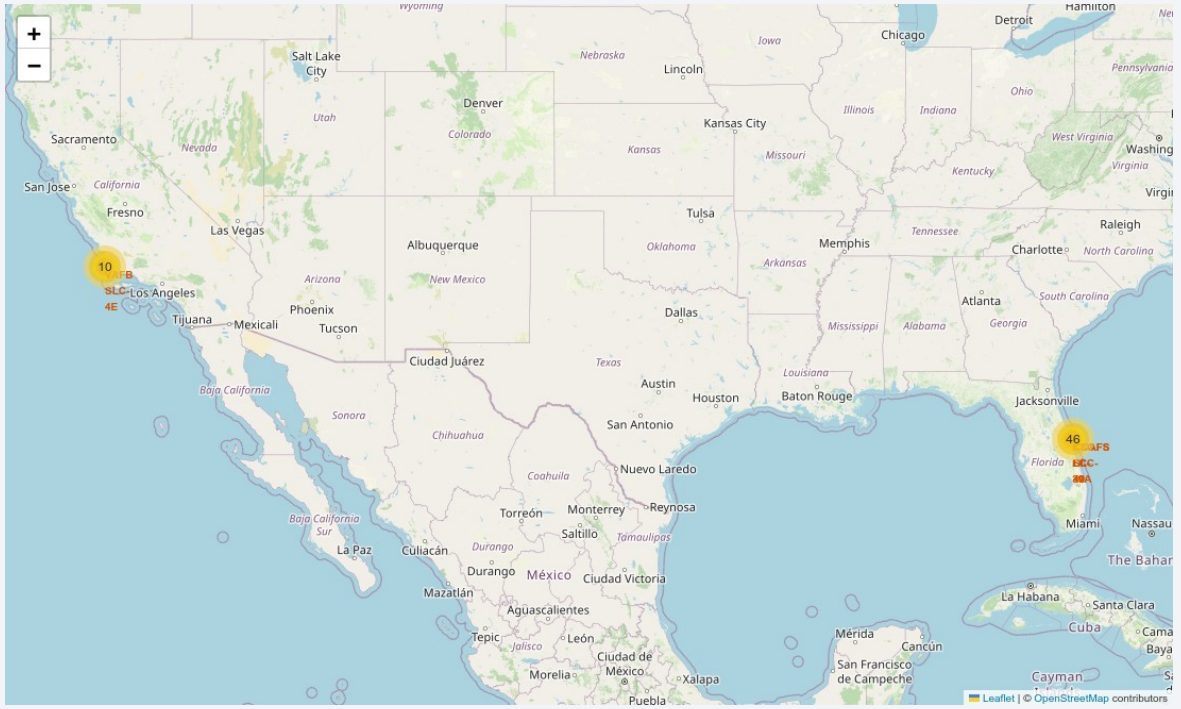
General questions:

Are all launch sites in proximity to the Equator line?

Yes, they are the closest possible locations on US mainland close to the Equator line.

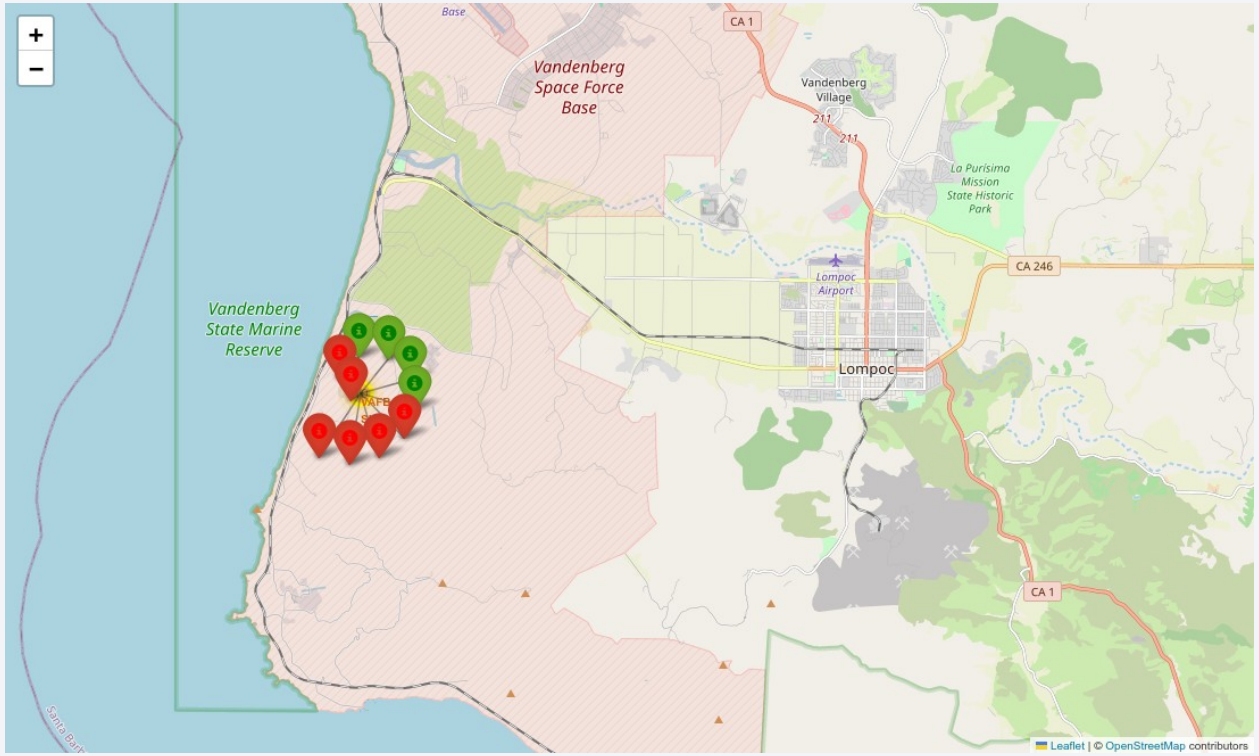
Are all launch sites in very close proximity to the coast?

Yes, they are.



Success/Failed Launches For Launch Site VAFB SLC-4E, California

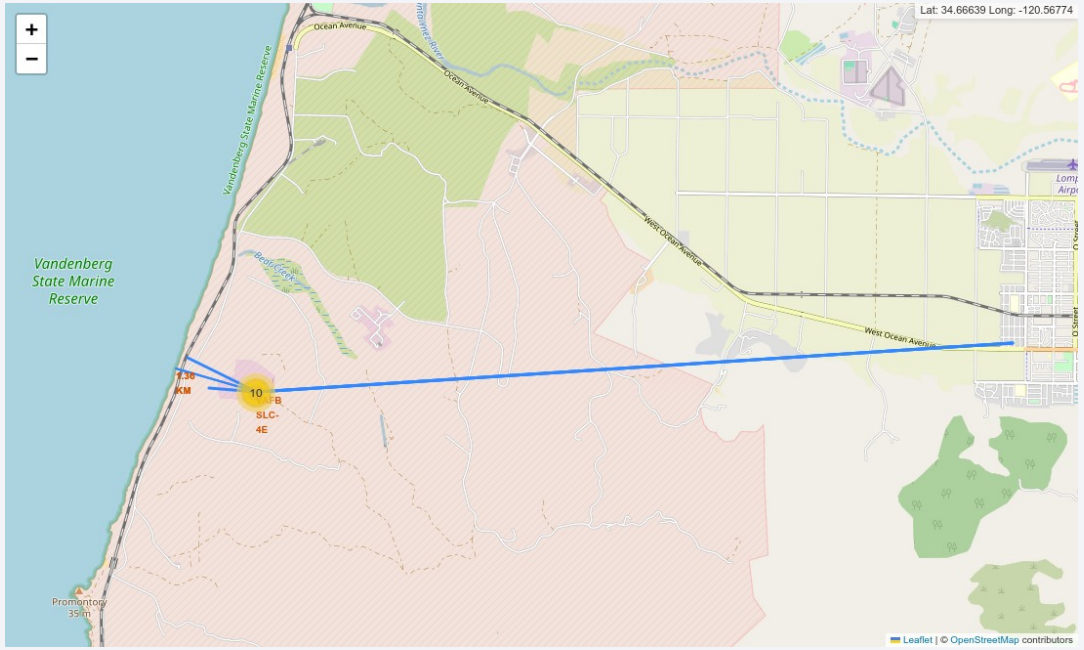
On the map of the VAFB SCC-4E Launch Site in California green markers show, if a launch was successful and red markers show, if the launch was a failure.

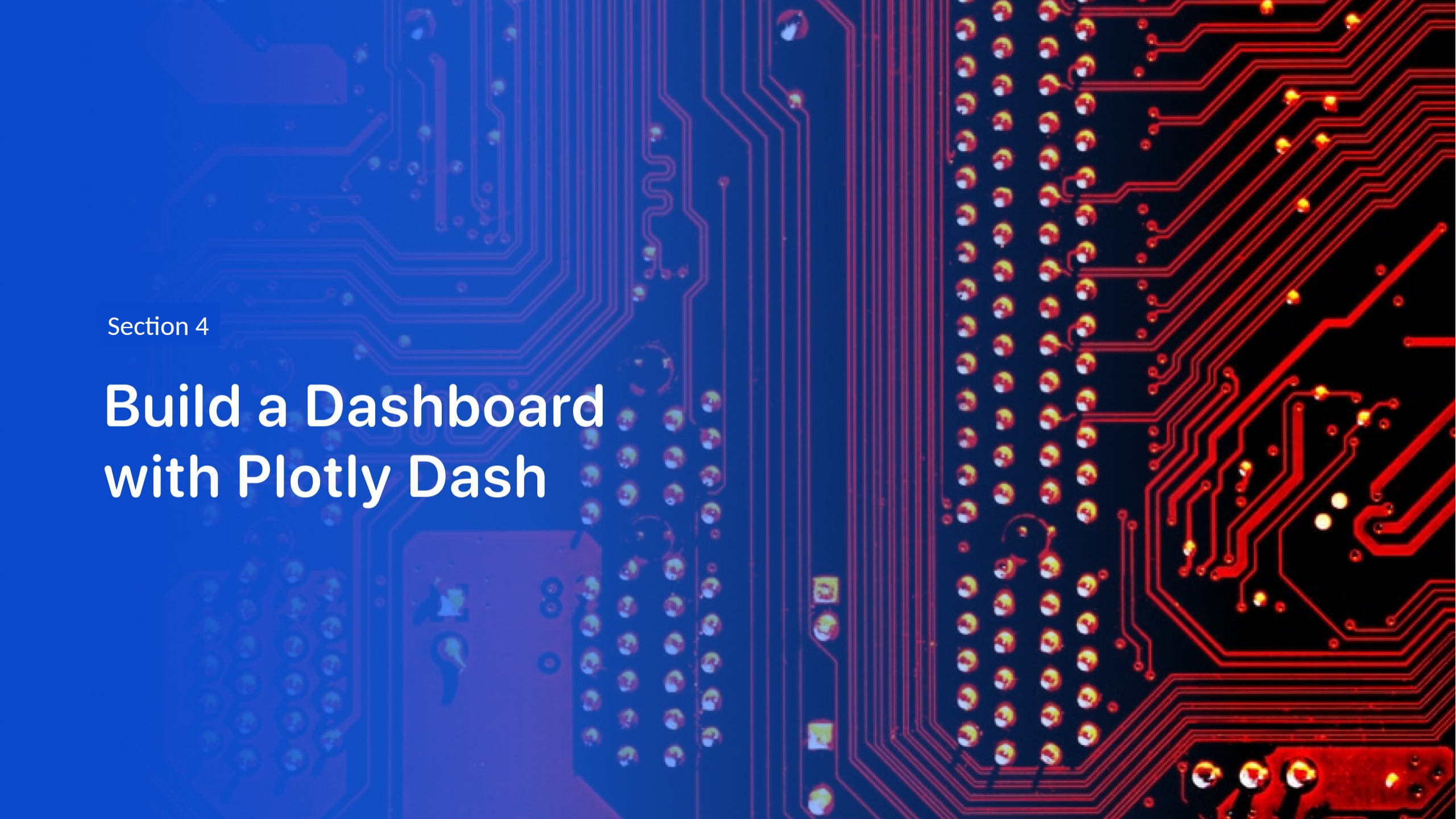


Distance between launch site VAFB SLC-4E to its proximities

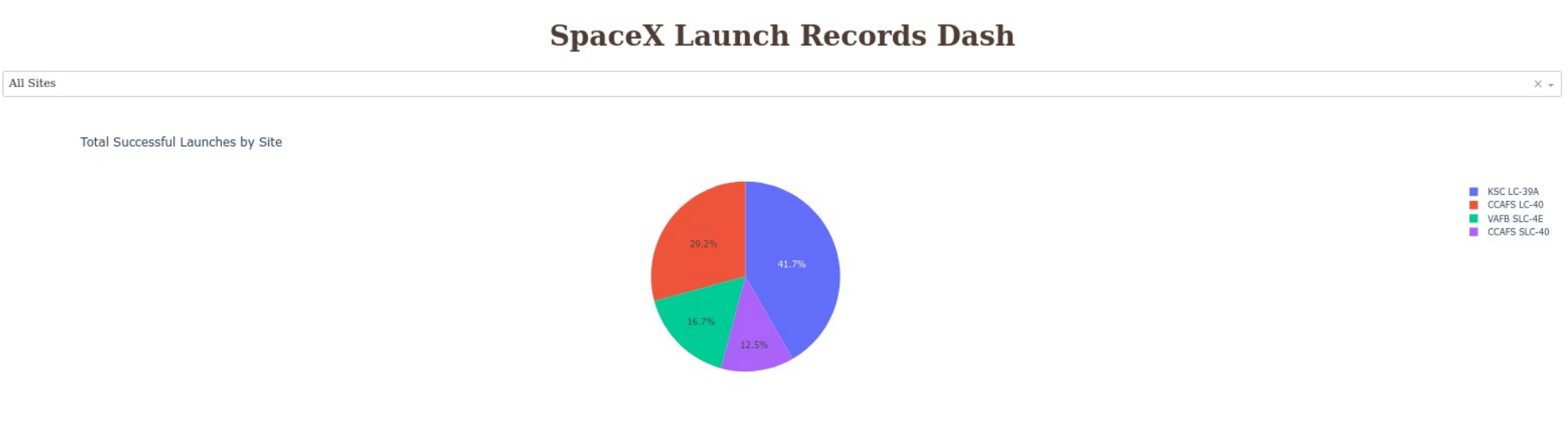
General Questions:

- Are launch sites in close proximity to railways? **yes**
- Are launch sites in close proximity to highways? **no**
- Are launch sites in close proximity to coastline? **yes**
- Do launch sites keep certain distance away from cities? **yes**



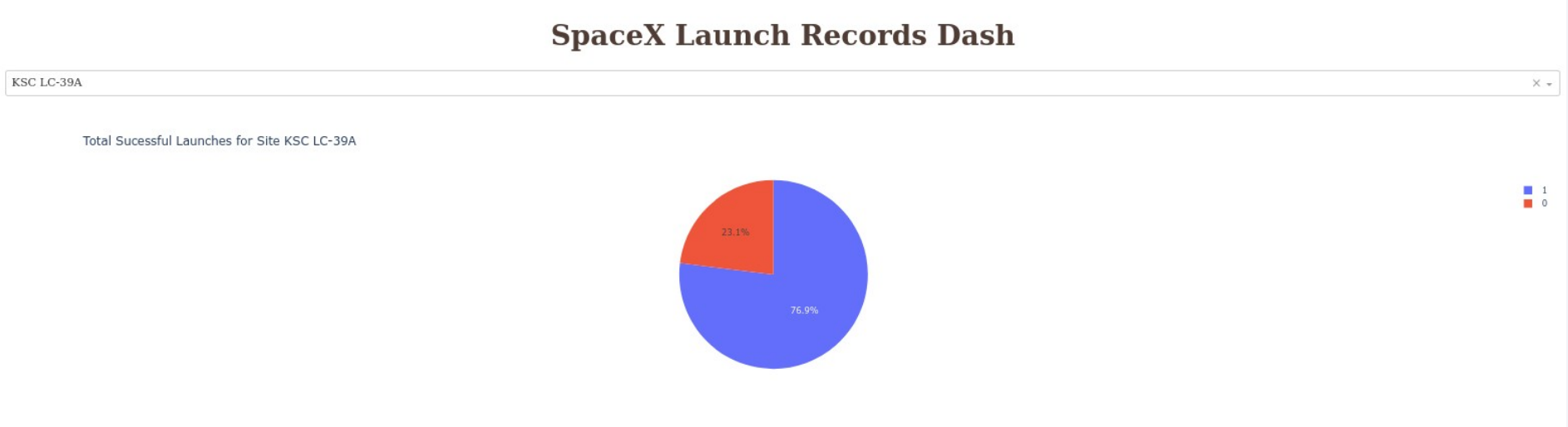

Section 4
Build a Dashboard with Plotly Dash

Launch success count for all sites



The chart shows that from all the sites, KSC LC-39A has the most successful launches.

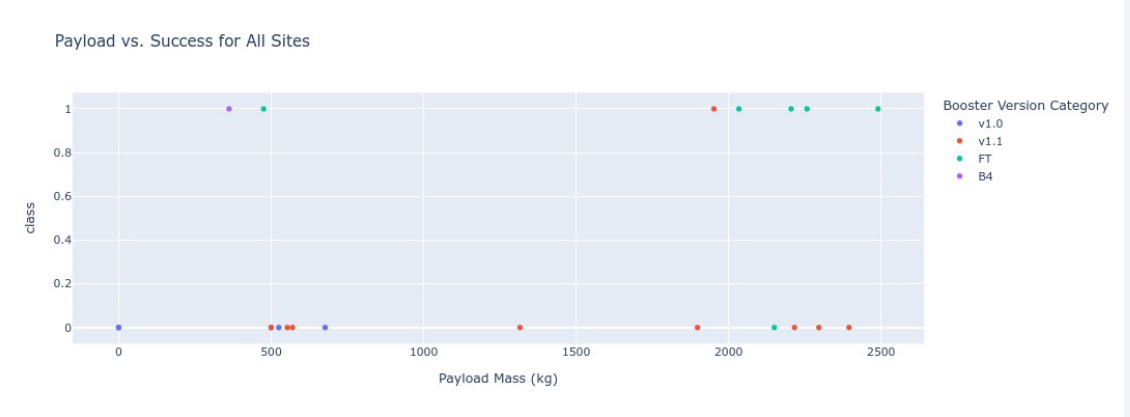
Launch site with highest launch success rate

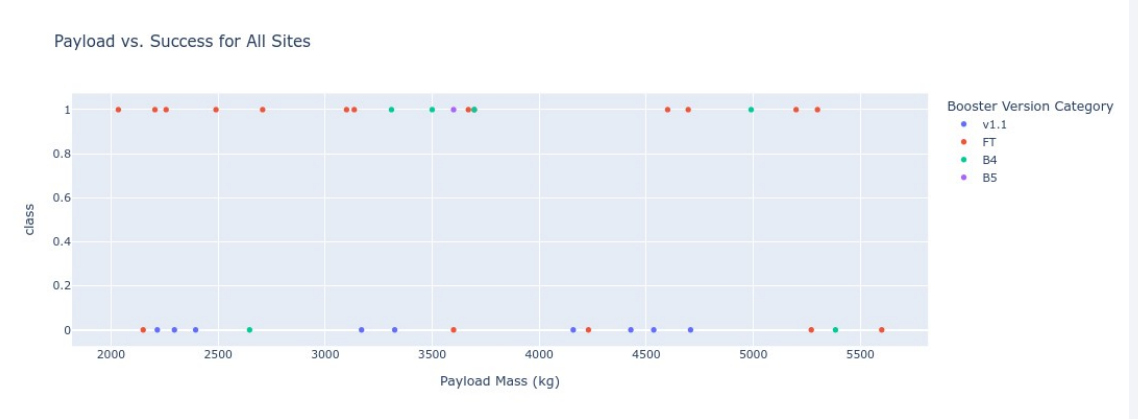


KSC LC-39A has the highest launch success rate (76.9%) with 10 successful launches.

Payload vs. Success for all sites

Payload range 2000 – 5500 kg has the highest launch success rate with 17 successful launches





Section 5

Predictive Analysis (Classification)

Classification Accuracy

Logistic Regression:
Test accuracy:83.33%

SVC:
Test accuracy:83.33%

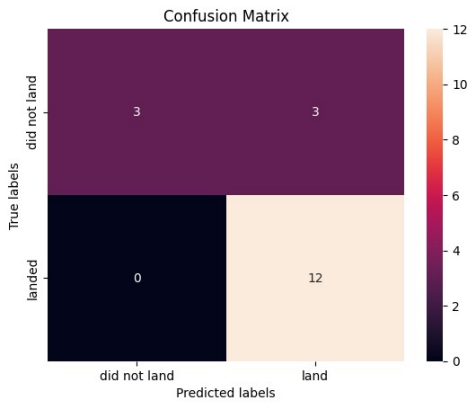
Decision tree classifier:
Test accuracy:83.33%

KNN:
Test accuracy:83.33%

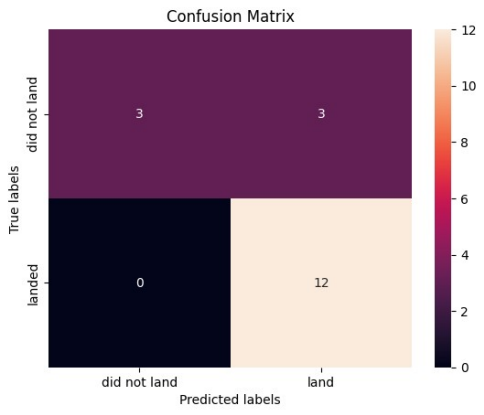
Basically, all methods perform equally good on the test data

Confusion Matrix

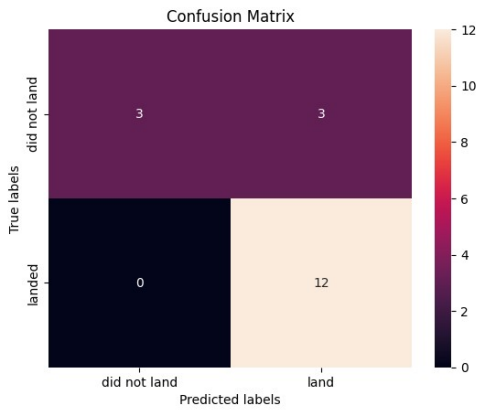
Logistic Regression:



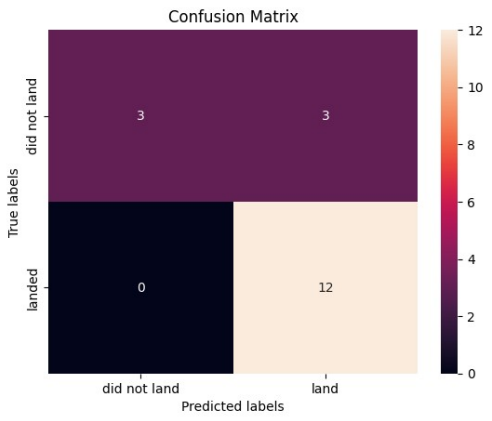
SVC:



Decision tree classifier:



KNN:



Conclusions

- Most of launch sites are in proximity to the Equator line and all the sites are in very close proximity to the coast.
- The success rate of launches increases over the years.
- KSC LC-39A has the highest success rate of the launches from all the sites.
- Orbits ES-L1, GEO, HEO and SSO have 100% success rate.
- Newer rocket technology and re-usability advancements contribute to increased success rates. Thus, the success rates of SPACEX launches is directly proportional to time, so in years they will eventually perfect the launches.
- Over all, there seem to be no clear correlation between payload mass and success rate
- Orbit type ISS & LEO: Higher payloads tend to succeed more often, while smaller payloads experience more failures
- Basically, the accuracy of all used modeling methods perform equally good

Appendix

https://github.com/jmw-geo/capstone

Thank you!