Winning Space Race with Data Science

Tessema Hirbaye
October 2024

Outline

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

Executive Summary

Summary of methodologies

Summary of all results

This project focuses on collecting and analyzing data from SpaceX to determine if first stage of Falcon 9 rocket launching success

The methodology employed include

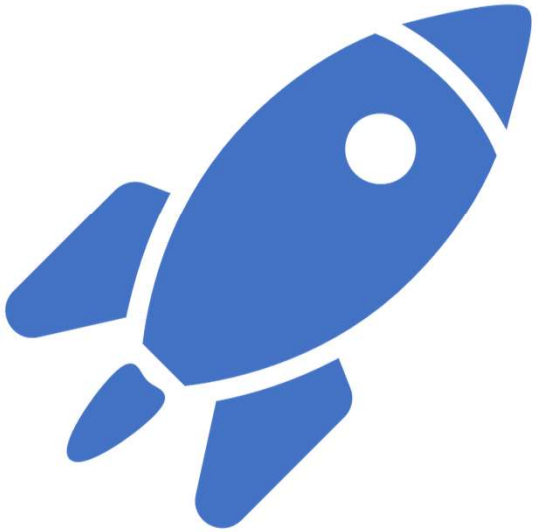
- Collecting SpaceX data using APIs and web scraping and data wrangling using API, filtering and dealing with nulls
- Perform EDA using SQL and visualization, interactive visual analytics and predictive analytics using classification models

Executive Summary

Results obtained shows that

- There are some relation or correlation between success rate and payload mass, orbit type and location of the launch site.
- Using machine learning classification models, first landing success is predicted for a new Company, Space Y that classification models showed 0.833 accuracy

Introduction



Now a days, commercial space companies working to make space travel affordable.

The most successful one is SpaceX while there is huge difference in cost of rocket launches between SpaceX and other providers

This project focuses on collecting and analyzing data from Space

The main objective of the project is

- To see if first stage of Falcon 9 rocket of SpaceX is successful
- Determine SpaceX reuses first stage so that we can determine a cost of launch
- Collect and analyze data from SpaceX

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

- Data was collected using APIs, from SpaceX Rest API and web scrapping

Perform data wrangling

- Data wrangling using API, normalize data, filtering some data and dealing with nulls

Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

- Preprocessing data, split data into training data and test data using Train_Test_Split
- Train the Models(Log.Regresion, SVM, D.Tree, and KNN), Test Models and plot Confusion matrix

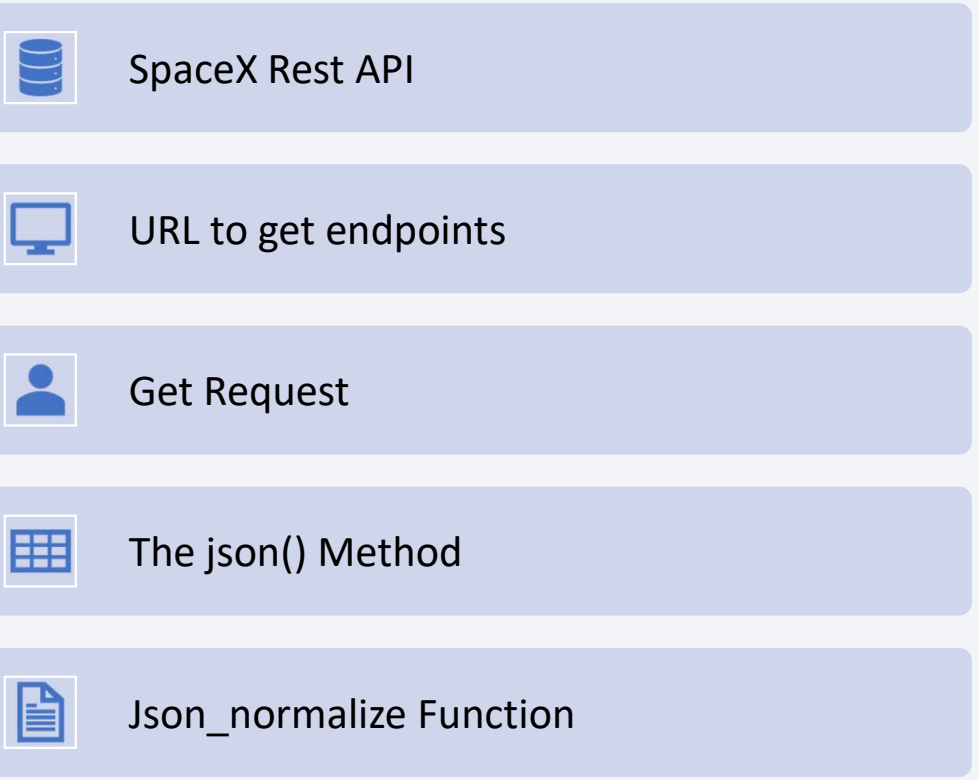
Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

- SpaceX Launch data was collected from API, SpaceX Rest API, using URL(api.spacexdata.com/v4/launches/past), different endpoints and get request to obtain launch data and parse the data.
- The data was convert into dataframe using normalize_json function.

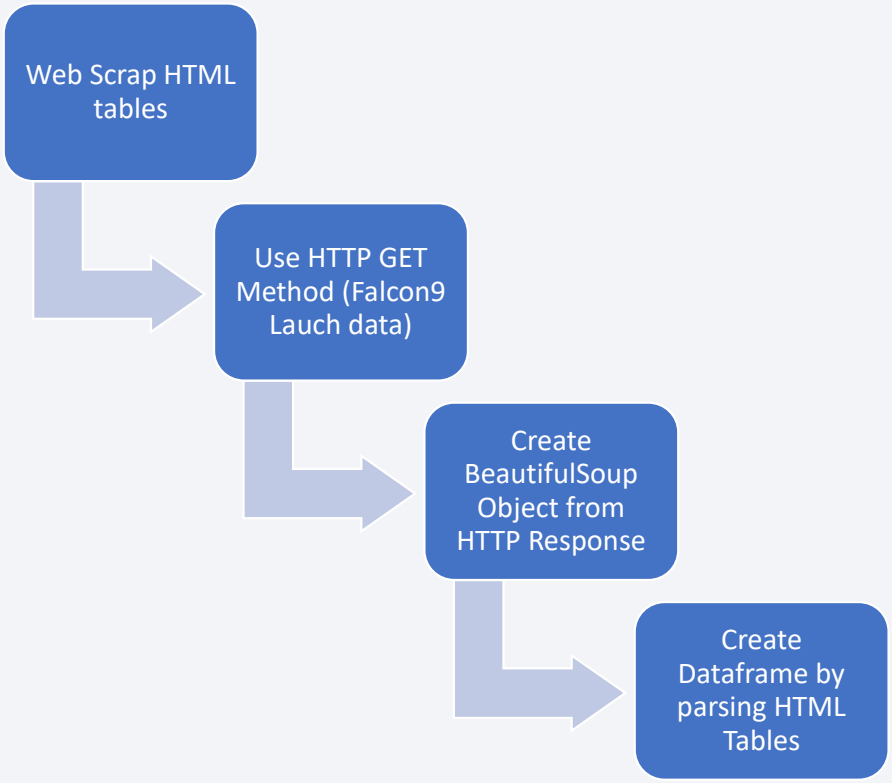
https://github.com/Tessemah24/Data-Science-Final-Project/blob/main/jupyter-labs-spacex-data-collection-api-v2.ipynb



Data Collection - Scraping

- HTML tables were web scrapped to collect Falcon9 Launch Records from Wikipedia using HTTP GET Method. BeautifulSoup object was created from response.
- Data from HTML tables was parsed and converted to pandas dataframe for visualization

https://github.com/Tessemah24/Data-Science-Final-Project/blob/main/jupyter-labs-webscraping%20(6).ipynb



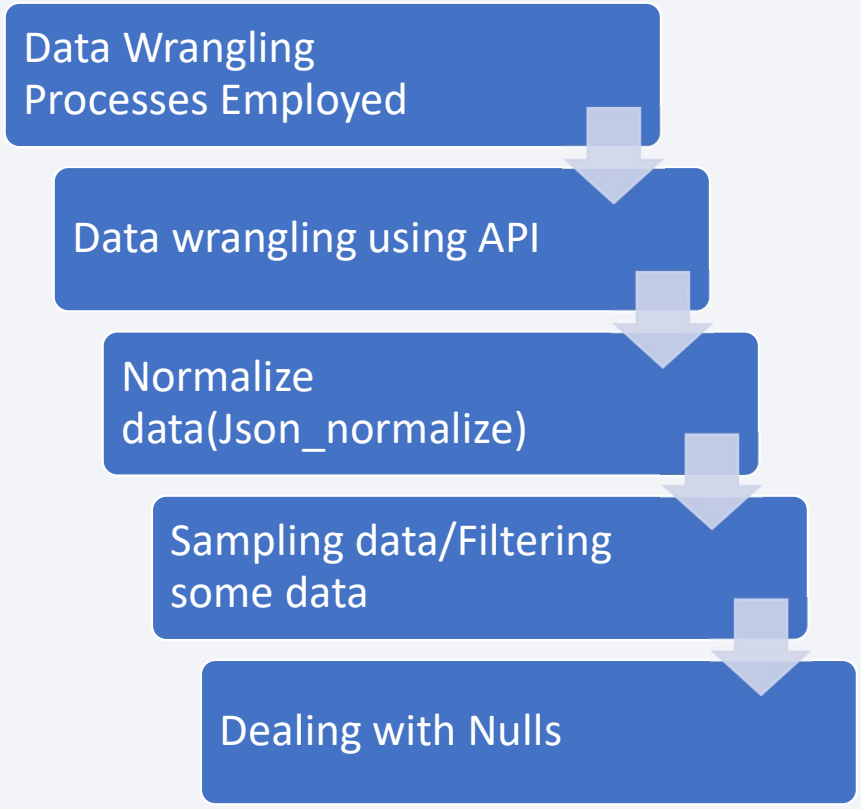
Data Wrangling

Data wrangling was done using API function, targets, and endpoints and exploratory data analysis to transforming data, deal with Nulls, filtering important data sets.

In addition, landing outcome was converted to classes to facilitate prediction.

- Data wrangling using API
- Normalize data(Json_normalize)
- Sampling data/Filtering some data
- Dealing with Nulls

https://github.com/Tessemah24/Data-Science-Final-Project/blob/main/labs-jupyter-spacex-Data%20wrangling-v2.ipynb



EDA with Data Visualization

- Exploratory data analysis was done using pandas and matplotlib and seaborn to visualize data using
  - Scatter point Chart (flight number vs Payload, Flight number Vs Launch Site, Payload Vs Launch Site) to see their effect on launch outcome.
  - Scatter Plot to see relationship between orbit type and success rate
  - Bar Chart to visualize which orbit type has highest success rate
  - Scatter plot to see relationship between orbit type vs Flight number with success and Payload vs Flight number with success
  - Line plot to see yearly success rate trend
- Dummy variables were created to categorial variables which are to be used for predictions
- https://github.com/Tessemah24/Data-Science-Final-Project/blob/main/jupyter-labs-eda-dataviz-v2.ipynb

EDA with SQL

- SQL Queries to do Exploratory Data Analysis (EDA)
  - **SELCET** statement, **LIMIT**, **ORDER BY** and **WHEWRE** statements
  - 'Alias' **AS and LIKE** operators
  - **MIN** and **MAX** functions and aggregations such as **SUM, COUNT, AVG**
- Established SQL extension (%load_ext SQL, con = sqlite3.connect("my_data1.db"), cur = con.cursor(), df.to_sql
- Query to select unique launch sites from spacextable(SELECT and DISTICT queries)
- Display 5 records, launch sites begin with the string 'CCA'(SELECT, WHERE, LIKE and LIMIT)
- Select and display Total Payload_mass in Kg booster launched by NASA(SELCET, SUM, AS, WHERE and LIKE)
- Display Average Payload mass (SELCET, AVG, WHERE)

- https://github.com/Tessemah24/Data-Science-Final-Project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Folium map was used to analyze launch site geo and proximities
  - By marking the launch site locations and their close proximities on an interactive map
  - Marking success/failed launches for each sit with green and red color markers
  - Distance between launch site and its proximities such as Railway, Coastline, and Highway were calculated
- Location Coordinates in Lat and Lon are as follows
  - CCAFS LC-40 28.562302, -80.577356
  - VAFB SLC-4E 34.632834, -120.610746
  - KSC LC-39A 28.573255, -80.646895
  - CCAFS SLC-40 28.561857, -80.577366
  - https://github.com/Tessemah24/Data-Science-Final-Project/blob/main/lab-jupyter-launch-site-location-v2.ipynb

Build a Dashboard with Plotly Dash

- The following plots/graphs and interactions were added to a dashboard.
- Launch Site Drop-down input component(to enable select one specific site among four sites to see the success count)
- Pie chart based on selected dropdown site using callback function to see total success rates and success and failure rates for each site
- Range Slider for Payload to see ranges of payload
- a callback function to render the success-payload-scatter-chart scatter plot

- https://github.com/Tessemah24/Data-Science-Final-Project/blob/main/Building%20Dashboard%20Application%20with%20Plotly%20Dash%202.ipynb

Predictive Analysis (Classification)

- Predictive Analysis was done using machine Learning Pipeline
- Preprocessing to standardize data
- Use Train_Test_Split to split data into training and Testing data
- Train the models(Logistic Regression, SVM, Decision Tree, KNN)
- Perform Grid Search to find hyperparameters that allow a model perform best
- Test each model (LR, SVM, DT classifier and KNN)
- Plot Confusion Matrix for each model
- https://github.com/Tessemah24/Data-Science-Final-Project/blob/main/SpaceX-Machine-Learning-Prediction-Part-5-v1%202.ipynb

Results

- Exploratory data analysis results
  - There appears relationship between Flight number and first stage landing outcomes
  - For VAFB-SLC launch site, there is no heavy payload mass greater than 1000
  - ES-LI, GEO, HEO, and SSO orbit types have higher success rate
  - Heavy Payload successful landing are related to PO, LEO and ISS Orbit types
  - Average Success rate increases Year 2013 to 2017 and it was stable in 2014
  - Total payload carried by boosters by NASA is 48213 kg
  - First successful landing outcome in ground pad was on 2015-12-22
  - There are 12 booster versions which carried the maximum payload mass

Results.....

Interactive analytics demo in screenshots

Folium Map Results

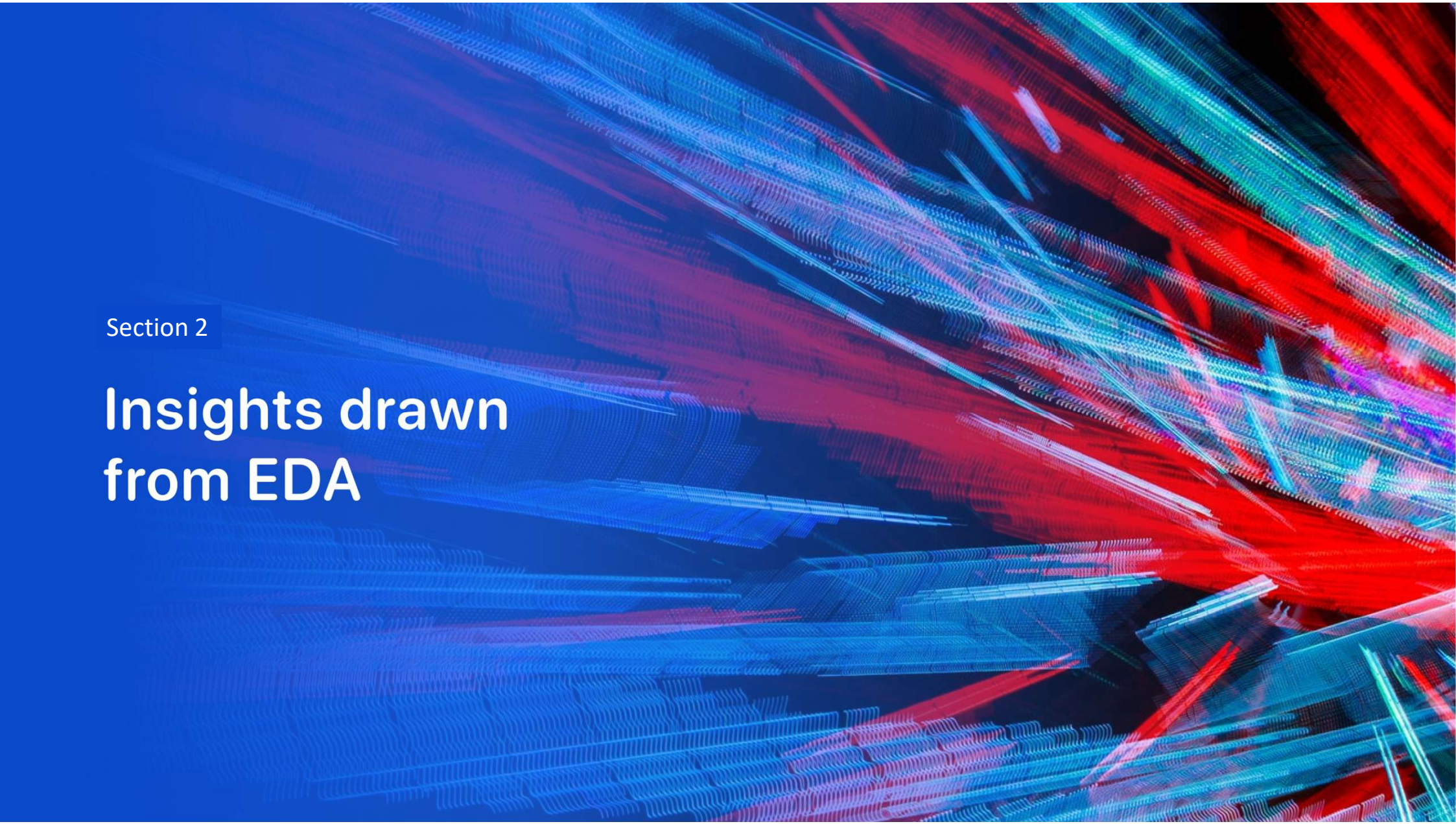
- All Launch Sites are located South-east and South-west costs of USA
- KSC LC -39A launch site has a highest success rate, 10 out of 13 are green(success)
- CCAFS SLC-40 has the lowest success rate, only 7 out of 26 are green
- Distances of launch site, CCAFS SLC-40 is 0.51 kms from coastline and 1.28 km from Railway/Train II Road

Dashboard Application with Plotly Results

- Pie-chart show that KSC LC-39A launch site has 41.7% success rate while CCAFS SLC-40 has only 12.5% success rate which is lowest
- Booster versions B4 and FT have highest success rate at 2k – 5k payload mass

Results

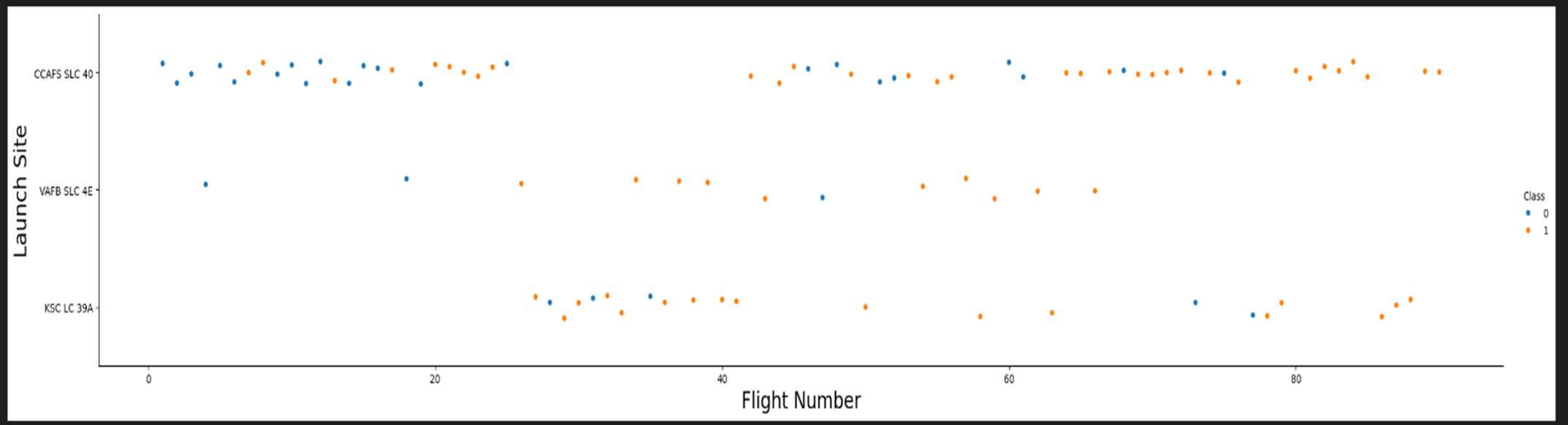
- **Predictive analysis results**
  - Logistic Regression has 0.0846 accuracy and 0.833 on test data using score
  - SVM has 0.848 accuracy and 0.833 on test data using score
  - Decision Tree classifier has 0.877 accuracy and 0.666 on test data using score
  - KNN Classifier has 0.848 accuracy and 0.833 on test data using score


Section 2
Insights drawn
from EDA

Flight Number vs. Launch Site

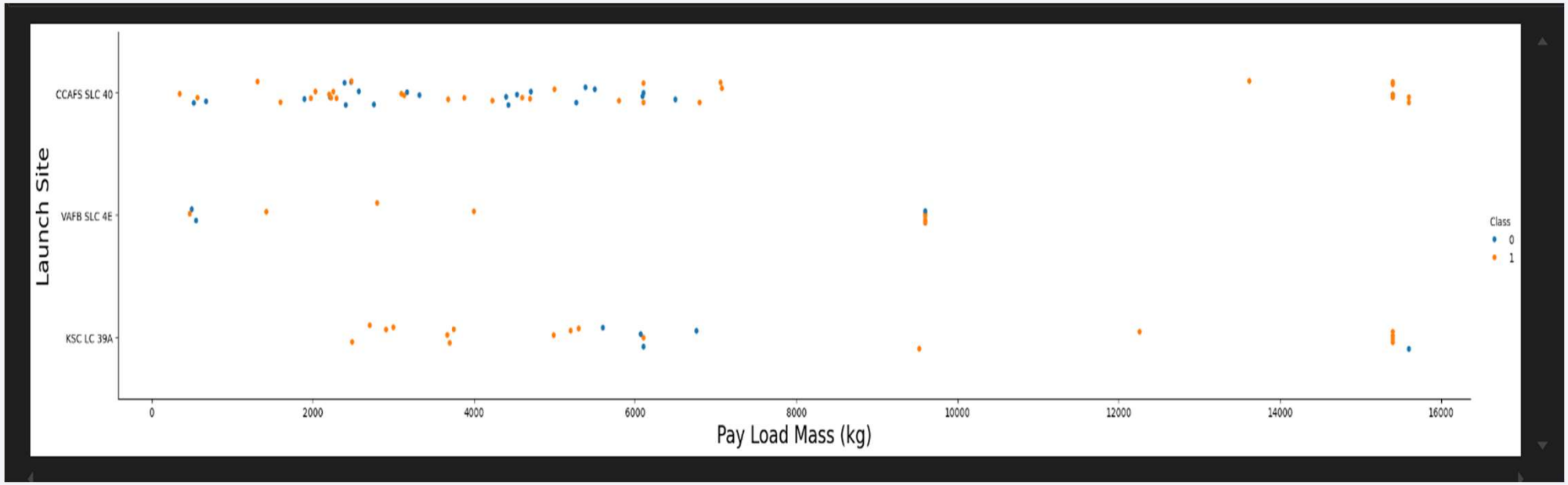
```
2  sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
3  plt.xlabel("Flight Number",fontsize=20)
4  plt.ylabel("Launch Site",fontsize=20)
5  plt.show()
```

✓ 0.1s



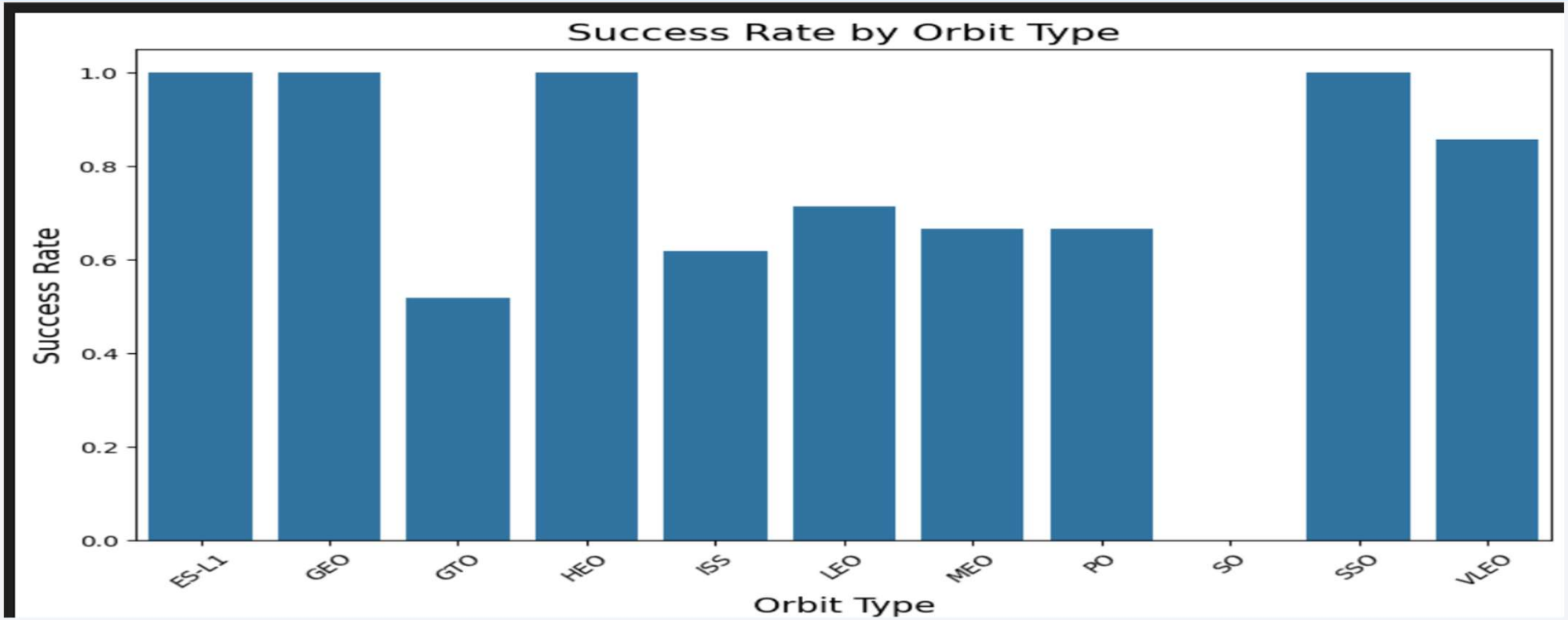
There is more success rate for flight numbers 80 and above for CCAFS SLC 40 and KSC LC 39A sites.

Payload vs. Launch Site



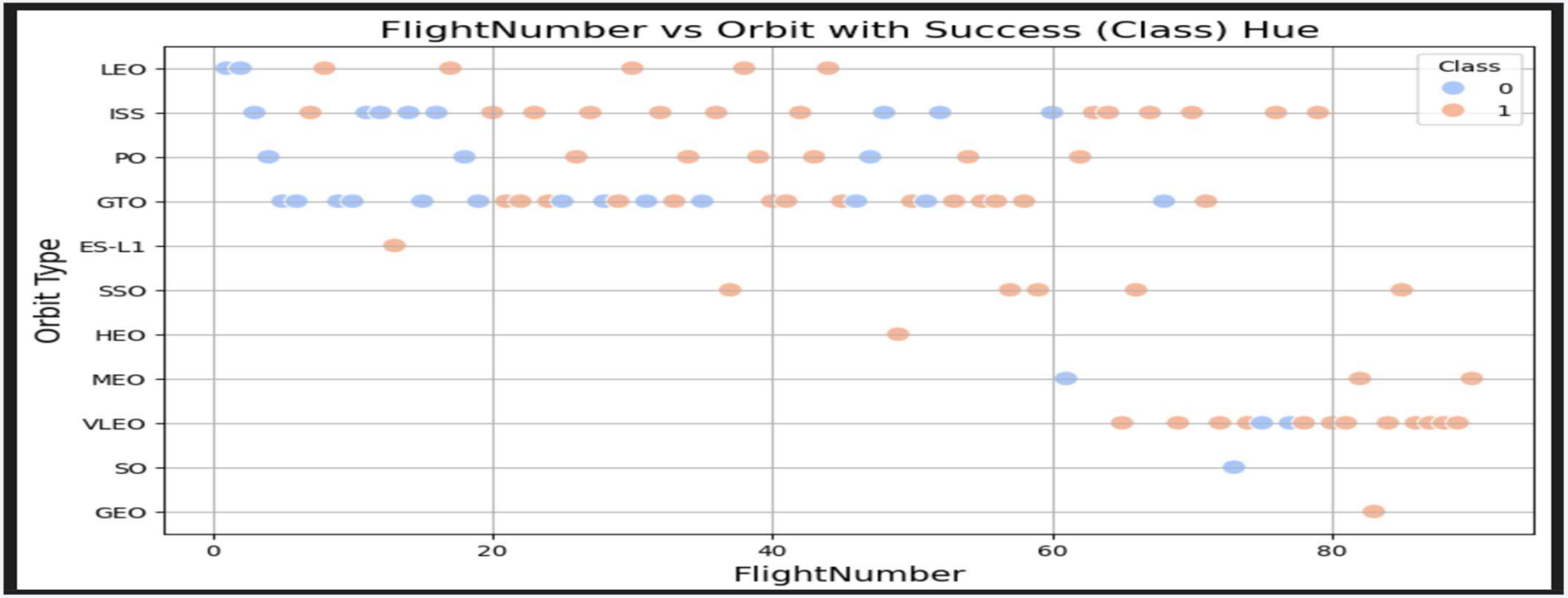
The scatter point plot shows that there is more success rate for KSC LC 39A site with low pay Load mass, < 400 and no rocket launched above 1000 pay load mass for VAFB-SLC 4E launch site.

Success Rate vs. Orbit Type



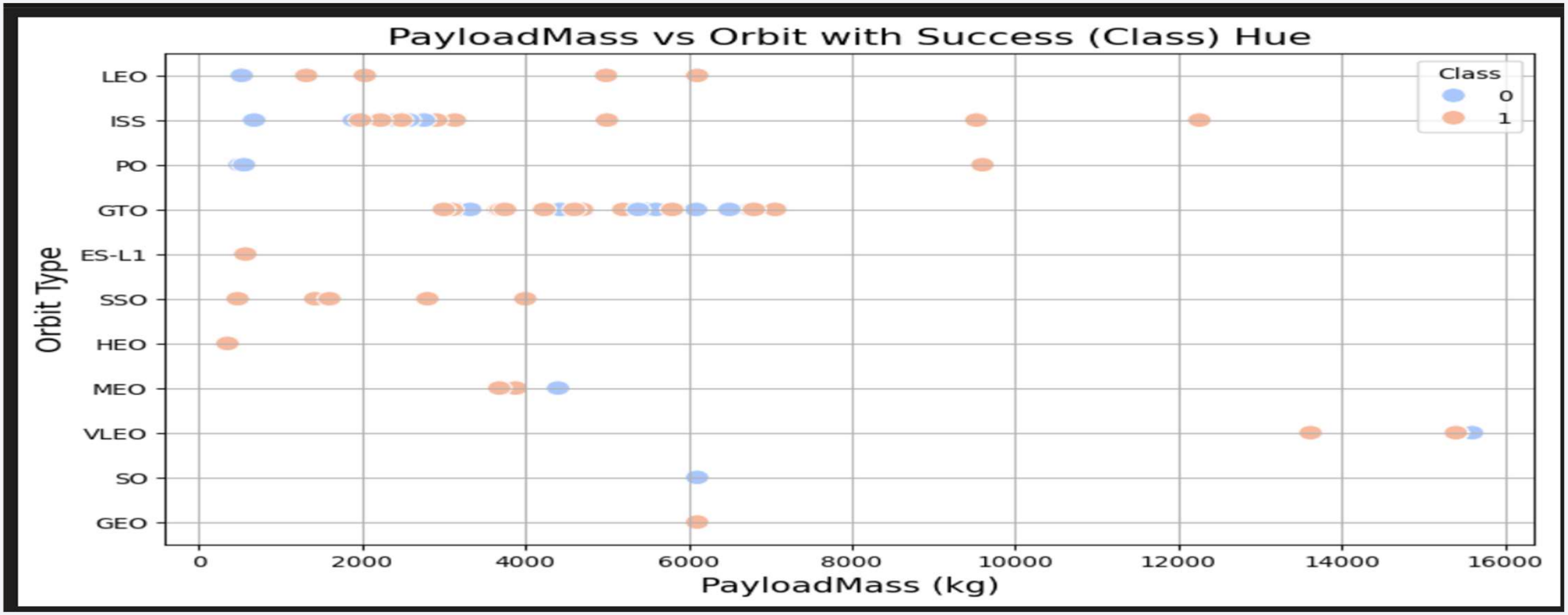
- The bar graph shows that ES-11, GEO, HEO, and SSO orbit types have highest success rate whereas SO orbit has zero success rate.

Flight Number vs. Orbit Type



Scatter plow shows that for LEO orbit the success relates to Flight number whereas there appears no relationship between flight numbers in GEO orbit.

Payload vs. Orbit Type



For heavy pay loads the successful landing are more for PO, LEO and ISS. However, for GTO orbit, it appears no relation.

Launch Success Yearly Trend

Line plot shows that launch success rate remain the same between 2010 and 2013 whereas it increases from year to year except between 2014 and 2015 where it remains the same. The trend decreases between 2017 and 2018.

All Launch Site Names

```
1  query = "SELECT DISTINCT Launch_Site FROM SPACEXTABLE"
2  unique_launch_sites = pd.read_sql(query, con)
3  unique_launch_sites
```

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

As it can be seen from screenshot of a query result above, there 4 unique launch sites in the space mission.

Launch Site Names Begin with 'CCA'

	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

As it can be seen from the above screenshot, 5 records were displayed using the query "SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5"

Total Payload Mass

```
query = "SELECT SUM(PAYLOAD_MASS__KG_) as Total_Payload_Mass FROM SPACEXTABLE WHERE Customer LIKE 'NASA (CRS)%'"
total_payload_mass = pd.read_sql(query, con)
total_payload_mass
```

	Total_Payload_Mass
0	48213

Total Payload Mass carried by Boosters Launched by NASA(CRS) is calculated using SQL query and it is **48213 kg.**

Average Payload Mass by F9 v1.1

```
query = "SELECT AVG(PAYLOAD_MASS__KG_) as Avg_Payload_Mass FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1'"
avg_payload_mass = pd.read_sql(query, con)
avg_payload_mass
```

	Avg_Payload_Mass
0	2928.4

Average Payload mass carried by Booster Version F9 V1.1 is calculated using SQL Query and it is **2928.4 Kg.**

First Successful Ground Landing Date

```
query = "SELECT MIN(Date) as First_Successful_Landing FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'"
first_success_landing = pd.read_sql(query, con)
first_success_landing
```

	First_Successful_Landing
0	2015-12-22

The first successful landing outcome in the ground was achieved on 22 December 2015 as per the above query result.

Successful Drone Ship Landing with Payload between 4000 and 6000

	Customer
0	SKY Perfect JSAT Group
1	SKY Perfect JSAT Group
2	SES
3	SES EchoStar

Names of the boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are listed on screenshot.

Total Number of Successful and Failure Mission Outcomes

```
query = """
SELECT Landing_Outcome, COUNT(*) as Total_Missions
FROM SPACEXTABLE
GROUP BY Landing_Outcome
"""
mission_outcomes = pd.read_sql(query, con)
mission_outcomes
```

	Landing_Outcome	Total_Missions
0	Controlled (ocean)	5
1	Failure	3
2	Failure (drone ship)	5
3	Failure (parachute)	2
4	No attempt	21
5	No attempt	1
6	Precluded (drone ship)	1
7	Success	38
8	Success (drone ship)	14
9	Success (ground pad)	9
10	Uncontrolled (ocean)	2

- Total number of successful and failure mission outcomes are presented as shown on screenshot of a SQL query for each type of landing outcome.

Boosters Carried Maximum Payload

```
query = """
SELECT Booster_Version
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE
)
"""
max_payload_booster = pd.read_sql(query, con)
max_payload_booster
```

	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

Here the names of the Booster Versions which have carried the maximum payload mass are presented as can been seen from SQL query result. There are 12 Booster versions.

2015 Launch Records

```
query = """
SELECT substr(Date, 6, 2) as Month, Booster_Version, Launch_Site, Landing_Outcome
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Failure (drone ship)'
AND substr(Date, 0, 5) = '2015'
"""
failure_2015 = pd.read_sql(query, con)
failure_2015
```

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

As it can bee seen from the screenshot of SQL query results, F9v1.1B1012 and F9v1.1B1015 are Booster versions failed on drone ship on launch site CCAFS LC-40 in the year 2015.

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

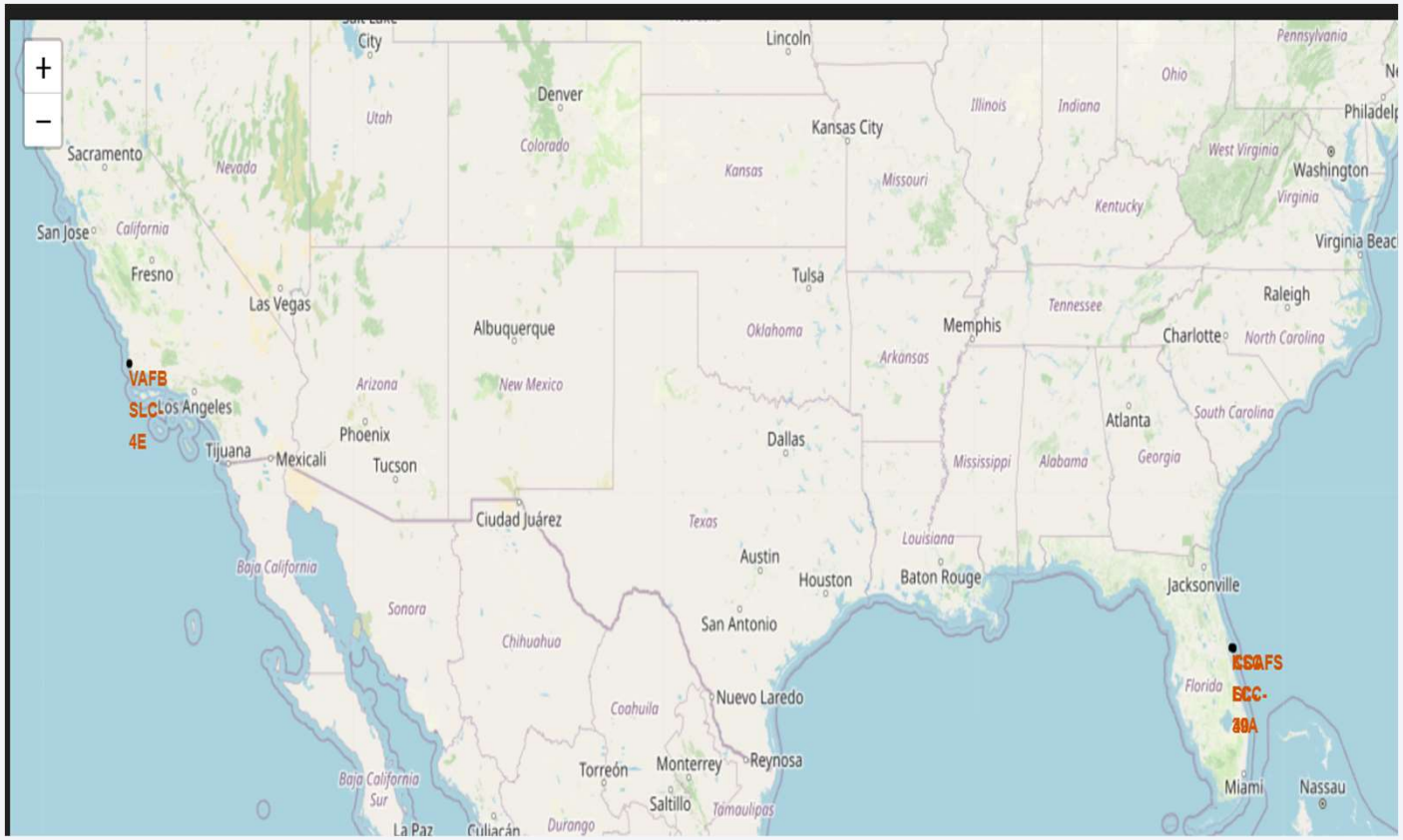
```
query = """
SELECT Landing_Outcome, COUNT(*) as Outcome_Count
FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY Outcome_Count DESC
"""
landing_outcome_rank = pd.read_sql(query, con)
landing_outcome_rank

```

	Landing_Outcome	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

As can be seen from the screenshot of a query results, landing outcomes between the date 2010-06-04 and 2017-03-20 are presented in the order of their outcome count.


Section 3
Launch Sites
Proximities Analysis

Folium Map of Launch Sites.



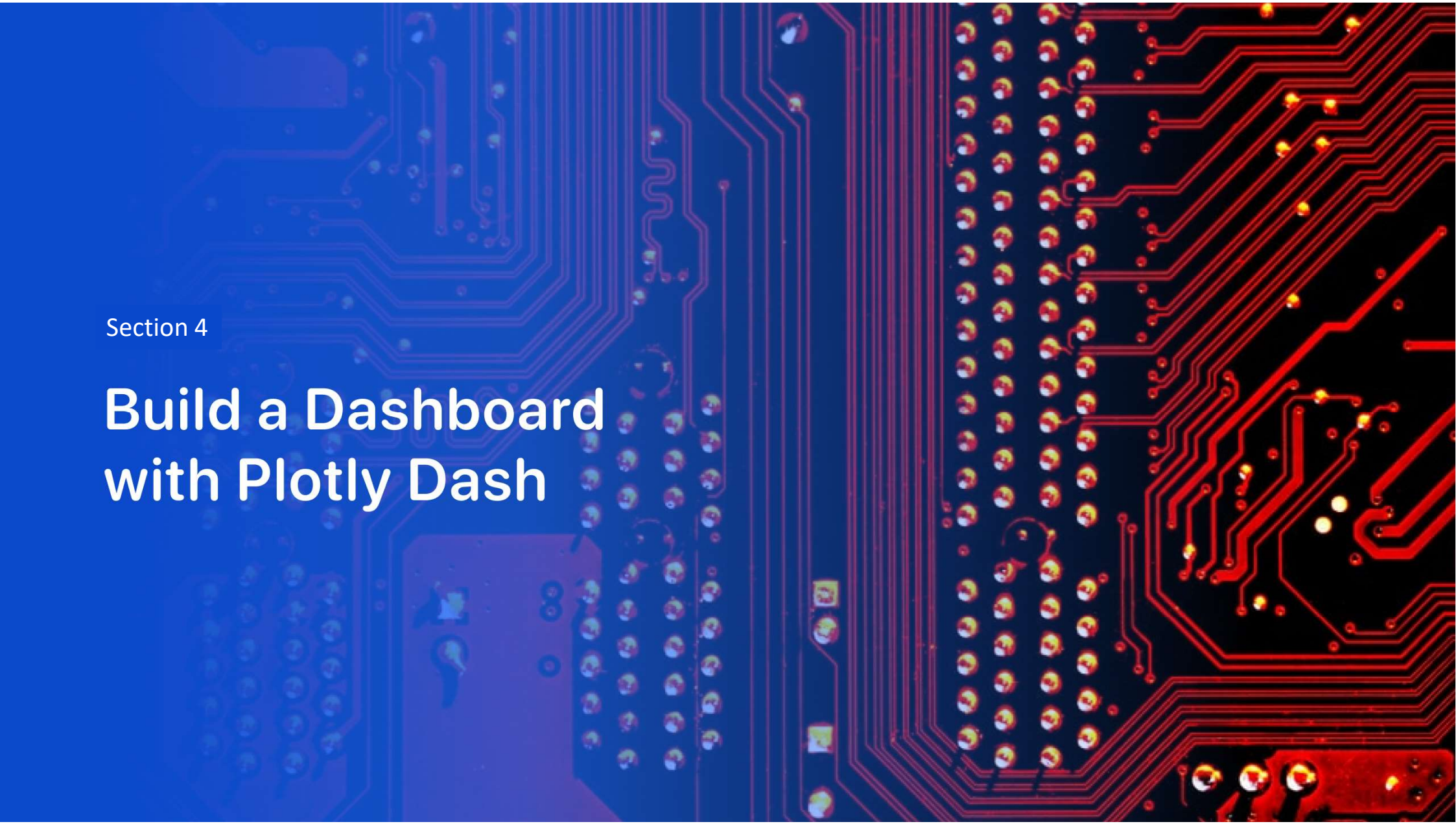
- Folium map showing launch sites marked in red, KSC LC-39 and CCFAFs-LC 40 located at South-east cost and VAFB SLC-4E located South-West cost in USA.

Folium Map success/failed launches for each site.

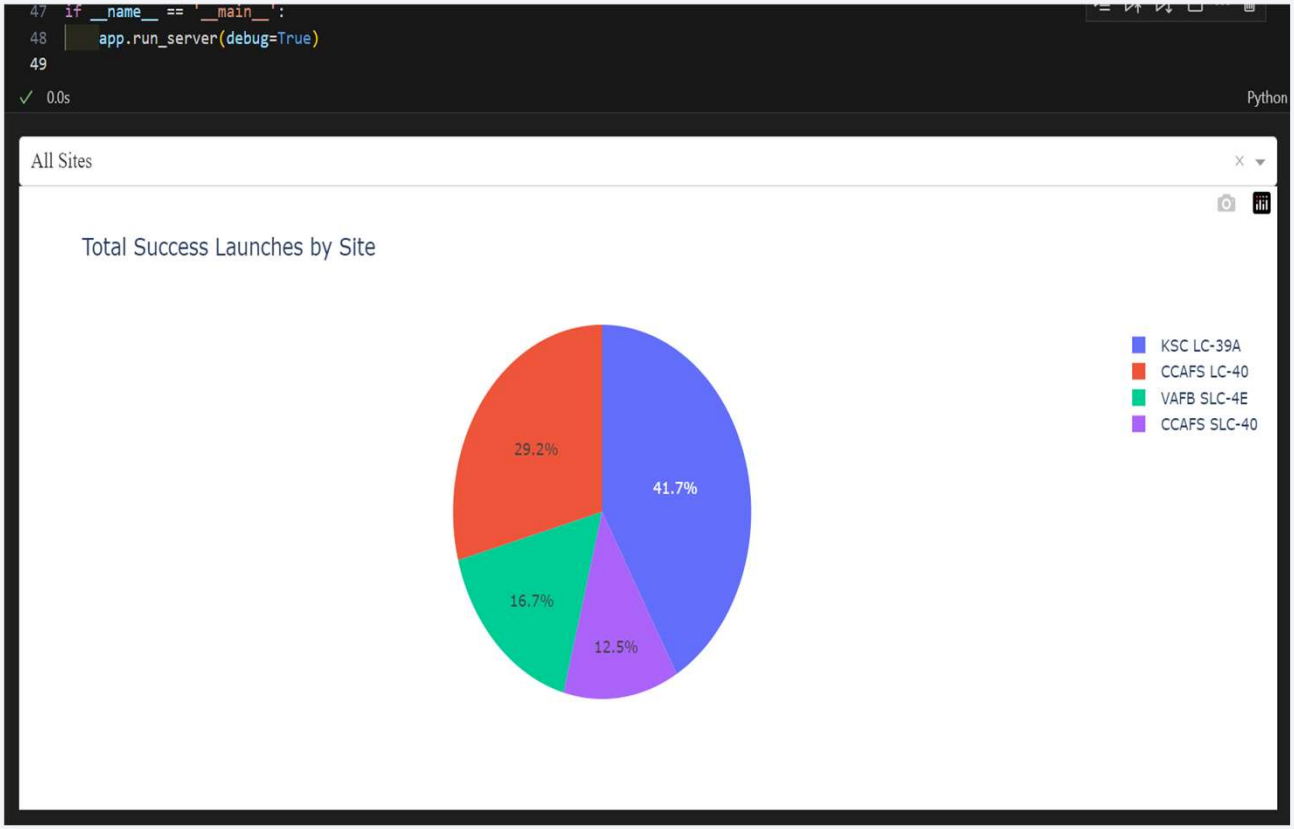
Screenshot of Folium Map shows low success rate(green marker show success class and red marker shows failure class)

Folium Map Showing Launch Distances B/N Launch Site Proximities

- Folium Map showing distances between launch Site, CCAFS SLC-40 and its proximities such as highway, railway and coastline.

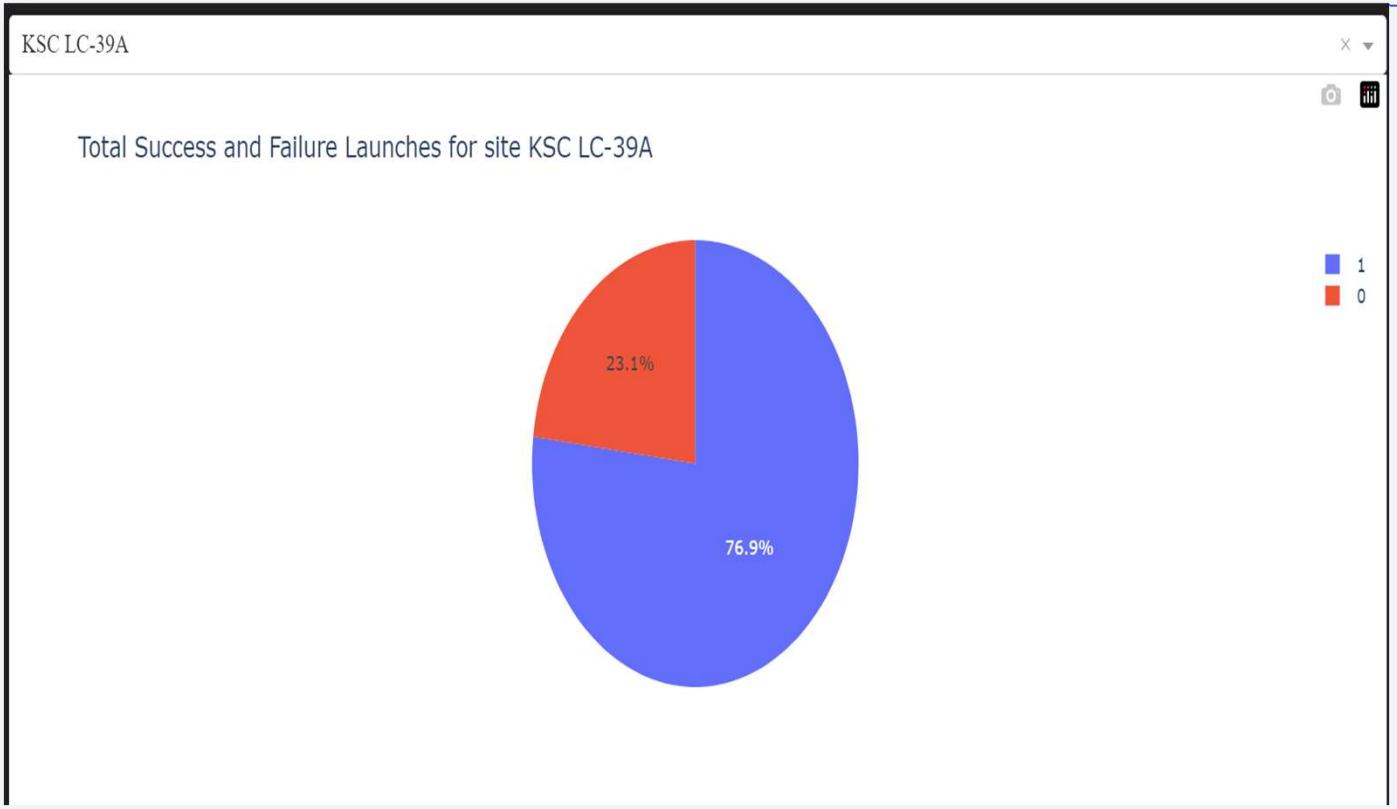

Section 4
Build a Dashboard with Plotly Dash

Dashboard Total Success Launches by Site



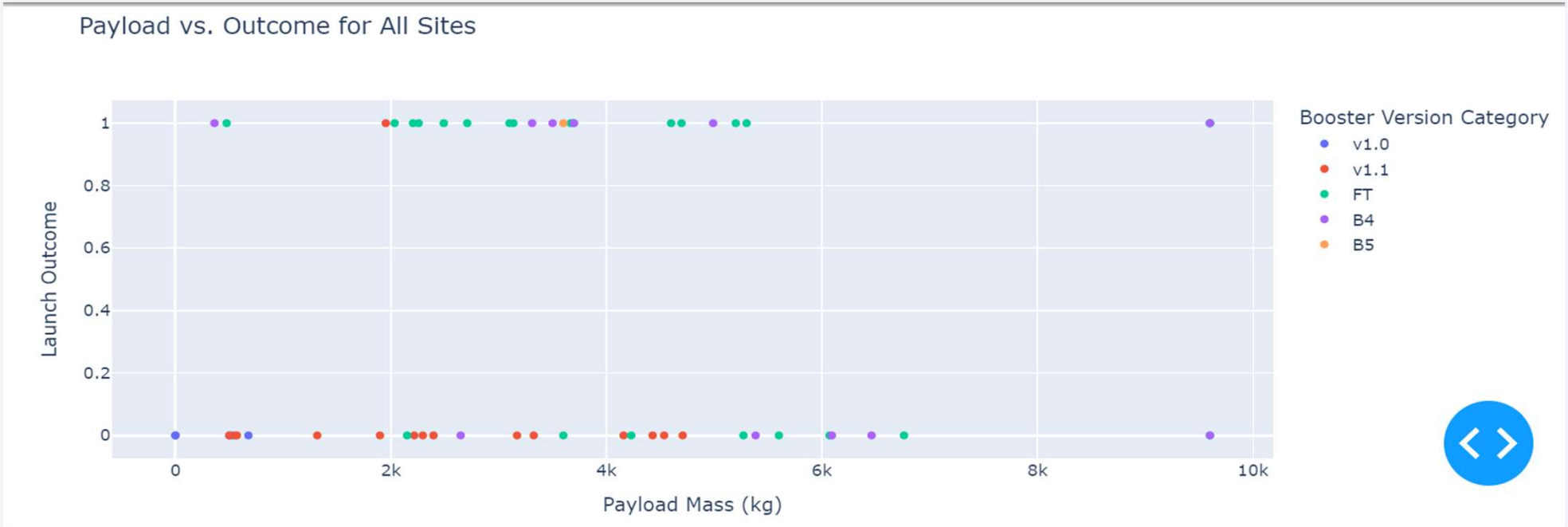
Pie Chart Showing total success launches by site. KSC LC-39A launch site has highest success rate which is 41.7% whereas CCAFS SLC-40 site has lowest success rate, 12.5%.

Dashboard: Total Success and Failure Launches for Site KSC LC-39A



Pie Chart shows 76.9% success launches for site KSC LC-39A which is the highest success launch among the four sites.

Payload Vs. Outcome for All Sites

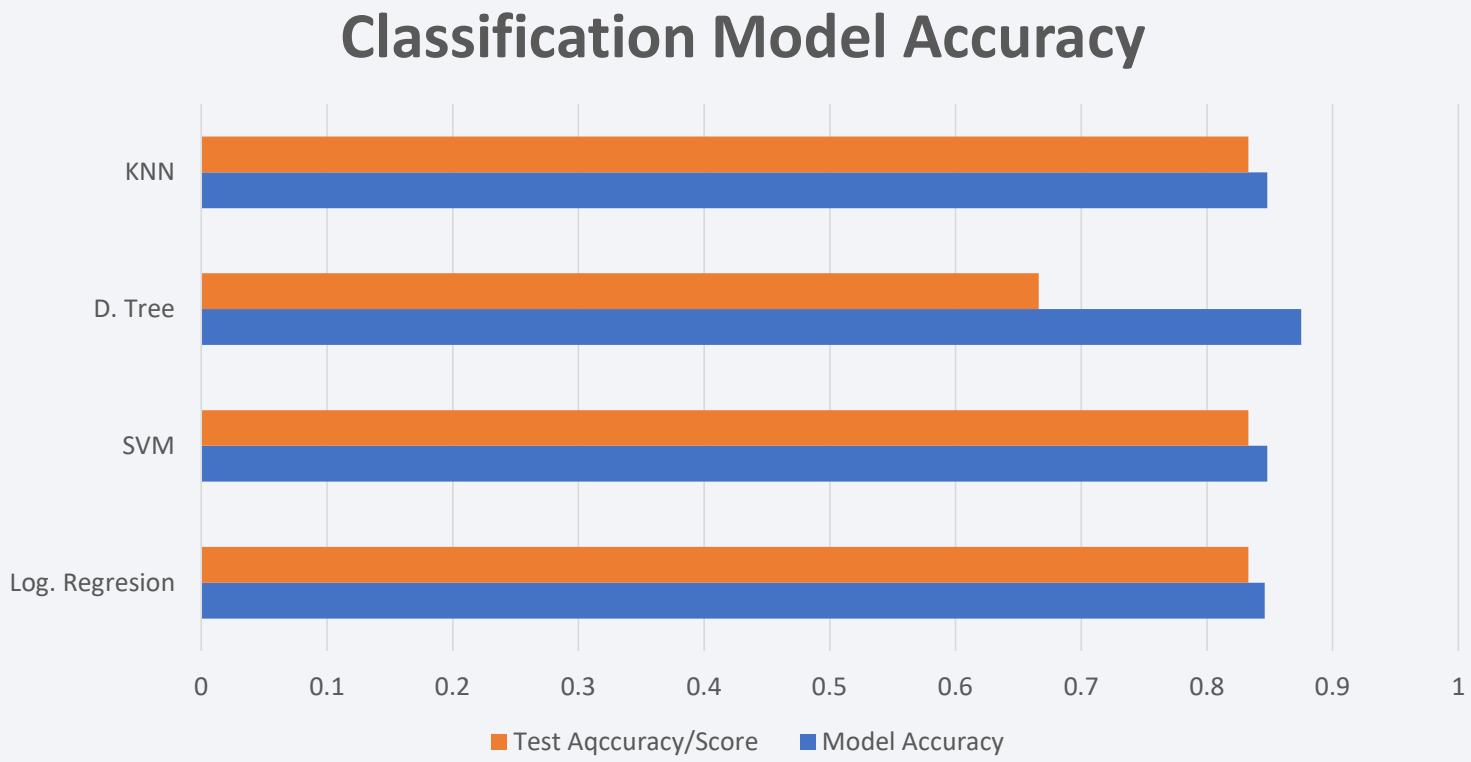


Scatter point plot shows that Booster Version Category FT and B4 have larger launch outcome for payload mass between 2k and 5k.

Section 5

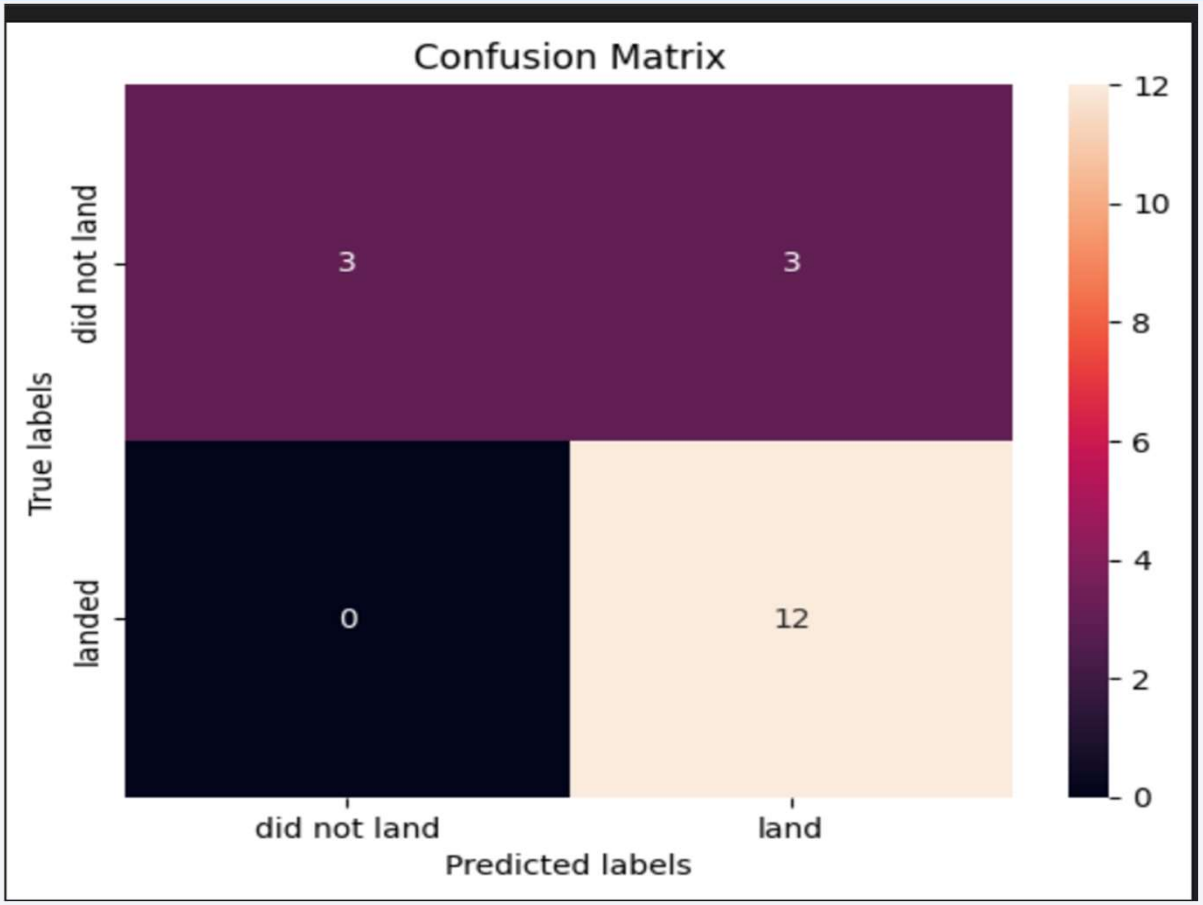
Predictive Analysis (Classification)

Classification Accuracy



- Log. Regression, SVM and KNN have the same classification accuracy on test data set on score, which is 0.833 and D. Tree classifier has lowest classification accuracy on test data.

Confusion Matrix for Log.Reg, SVM, and KNN



- The confusion matrix for Logistic Regression, Support Vector Machine(SVM) and KNN as shown on screenshot,
- True Positive – 12 true label is landed (predicted label is also landed
- False positive – 3 (True label not landed, predicted label is landed)

Conclusions

In EDA scatter plots show that success rate varies based on features such as Flight number, orbit type and payload mass while there is increasing trend in average yearly success rate.

As EDA using SQL shows, both payload capacity and landing success showed improvement with booster versions capable of carrying substantial payloads.

Large number of the booster version that carried the maximum payload mass highlights SpaceX's ability to transport heavy payloads,

Ranking of landing outcomes between 2010 and 2017, if failures dominated earlier in the period but success increased toward the later years reflecting SpaceX's technological progress in reusable rocket technology.

Conclusions

Marking a successful and failed launches on the map using green and red markers, provides an intuitive way to launch performance at different sites

For example, KSC LC 39A site has a higher concentration of green (successful) markers, while other 3 sites show a greater number of red (failed) markers

This can highlight trends in launch success rates based on geographical factors, site management, or other operational conditions.

Proximity of launch sites to the essential infrastructure such as railway, highways or coastlines highlights the critical infrastructure surrounding launch sites at South-east cost.

Conclusions

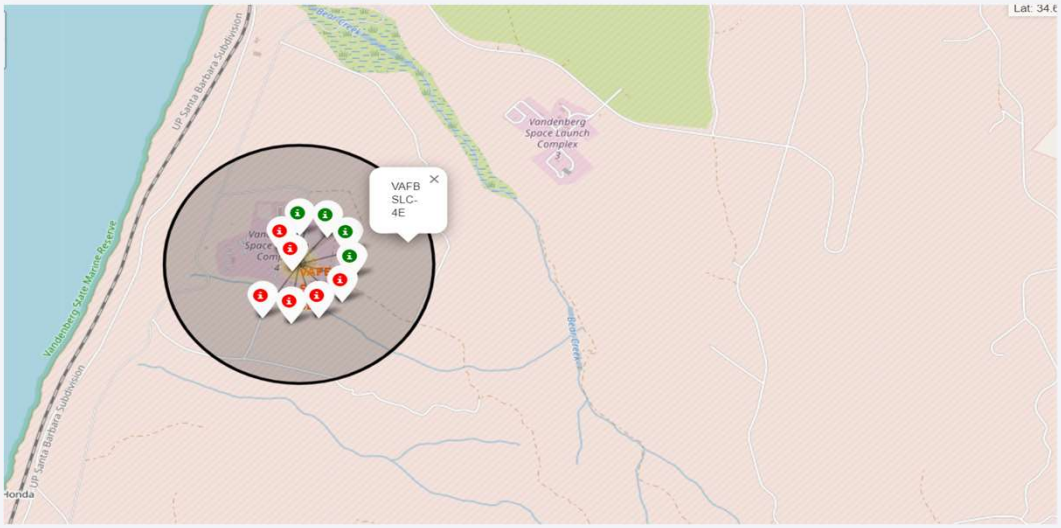
Different launch sites have varying success rates. By analyzing the pie chart, it is possible to identify the most reliable launch site for successful missions seems KSC LC-39A, 41.7% of overall success and 76.9% success rate of its own launches.

There may not be a strong correlation between payload mass and launch success, but certain payload ranges, example Booster Version Category FT and B4 have larger launch outcome for payload mass between 2k and 5k.

From EDA, interactive analytics and classification models, it appears that the launch success rate depend on payload mass, orbit type, location and proximities of the launch site .

Using machine learning classification models first landing for a new Company, Space Y is predicted with accuracy of 0.833 on test data

Appendix



Annex 1. Location of VAFB LC-4E launch site at Southwest Cost of USA

Thank you!