IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Alvaro Martinez Lacasta>
<May 2025>

Outline

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

Executive Summary

•**Goal:** Analyse SpaceX launches to understand success drivers and build a launch-success classification model.

•**Data sources:**

- SpaceX REST API (launch & landing data)
- Web-scraped tables for launch sites and boosters

•**Key findings:**

- Overall launch-success rate ≈ 90 %
- Highest success orbits: GTO & LEO (> 95 %)
- Best model: Random Forest (≈ 85 % accuracy)

•**Recommendations:**

- Add geospatial features to improve the model
- Ingest a longer landing-history series

Introduction

•**Context:** SpaceX pursues rocket reuse to cut costs and increase launch cadence.

Problem statement:

- What are the key factors (e.g., payload mass, orbit type, launch site) that influence the successful landing of Falcon 9’s first stage?
- Can we develop a predictive model to accurately forecast landing success based on historical launch data?
- How does the geographical location of launch sites and their proximities affect launch outcomes?
- How can interactive spatial analysis and dashboards support better decisionmaking for launch planning and site selection?
- This project seeks to answer these questions through data analysis, visualization, and machine learning, providing insights to improve launch success predictions and reduce operational risks.

Section 1

Methodology

Methodology

- **Data Collection**
  - SpaceX REST API
  - Web scraping of complementary data

- **Data Wrangling**
  - Cleaning, merging, feature engineering

- **Exploratory Data Analysis**
  - Python visuals (Matplotlib/Seaborn)
  - SQL queries in SQLite
  - Interactive Folium maps
  - Plotly Dash dashboard

- **Predictive Analysis**
  - Classification models (LogReg, SVM, RF, KNN)
  - Hyper-parameter tuning & cross-validation

Data Collection

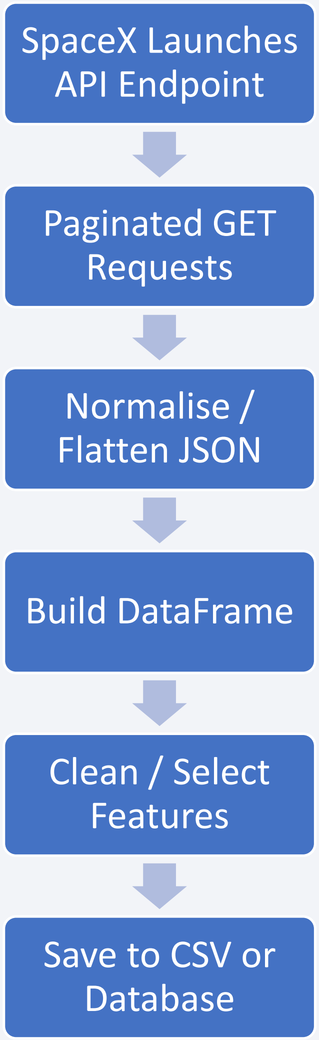
The data was collected using two methods:

- Using **get request to the SpaceX API**.
  - Next, we decoded the response content as a Json using .json() function call and turn it into a pandas dataframe using .json_normalize().

- In addition, we performed **web scraping from Wikipedia for Falcon 9** launch records with BeautifulSoup to collect more data.
  - The objective was to extract the launch records as HTML table, parse the table andconvert it to a pandas dataframe for future analysis.

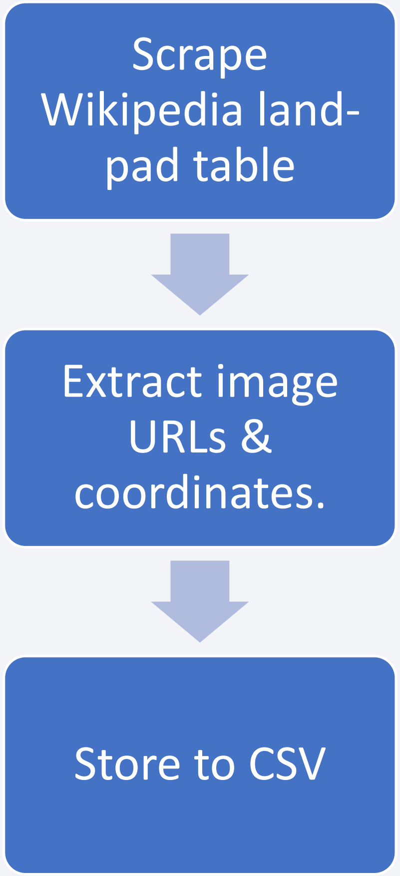
Data Collection – SpaceX API

- **Main endpoint**: https://api.spacexdata.com/v3/launches
- Used **get_request** to the SpaceX API
- Paginated calls for all launches.
    - Extract fields: flight_number, launch_site, payload_mass, orbit, launch_success, landing_success.
- **Normalize JSON** → DataFrame
- **GitHub link:** https://github.com/amlacasta/Data-Science-Fundae---IBM-SpaceY/blob/main/1%20jupyter-labs-spacex-data-collection-api%20v2.ipynb



Data Collection - Scraping

- **Goal:** Fetch booster details and infrastructure distances
- **Tools:** BeautifulSoup + Requests.
- **Steps:**
  - Scrape Wikipedia land-pad table.
  - Extract image URLs & coordinates.
  - Store to CSV
- **GitHub link:** https://github.com/amlacasta/Data-Science-Fundae---IBM-SpaceY/blob/main/2%20jupyter-labs-webscraping%20%20completo.ipynb



Data Wrangling

- **Goal:** Performed exploratory data analysis (EDA) and determined the training labels.
  - Calculated the number of launches at each site, and the number and occurrence of each orbits
  - Created landing outcome label from outcome column and exported the results to csv.

- **Merge**: API + scraped data on site_id
- **Transforms:**
  - Type conversions (dates, numerics)
  - Missing-value imputation (median payload)
  - Feature creation
- https://github.com/amlacasta/Data-Science-Fundae---IBM-SpaceY/blob/main/labs_jupyter_spacex_Data_wrangling_v2%20completo.ipynb

EDA with Data Visualization

Scatter Graphs plotted:

- Flight Number vs. Payload Mass
- Flight Number vs. Launch Site
- Payload vs. Launch Site
- Orbit vs. Flight Number
- Payload vs. Orbit Type
- Orbit vs. Payload Mas

Bar Graph being drawn:

- Mean (of class) vs. Orbit

Line Graph being drawn:

Success Rate vs. Year

Github: https://github.com/amlacasta/Data-Science-Fundae---IBM-SpaceY/blob/main/4%20jupyter-labs-eda-dataviz-v2.ipynb

EDA with SQL

Performed SQL queries to gather information about the dataset.

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'KSC’
- 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 where the successful landing outcome in drone ship was achieved.
- Listing the names of the boosters which have success in ground pad 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 records which will display the month names, successful landing_outcomes in ground pad, booster versions, launch_sites for the months in year 2017.
- Ranking the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

Build an Interactive Map with Folium

- Marked all launch sites, and added map objects such as markers, circles, lines to mark the success or failure of launches for each site on the folium map.
- Then assigned the feature launch outcomes to class 0 and 1.i.e., 0 for failure and 1 for success.
- Using the color-labeled marker clusters, identified which launch sites have relatively high success rate.
- Calculated the distances between a launch site to its proximities, to answer some question:

Example of some trends in which the Launch Site is situated in.

Are launch sites in close proximity to railways? No

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

Build a Dashboard with Plotly Dash

- Built an interactive dashboard with Plotly dash

- It takes input of Launch Site and Payload Mass from user

- Graphs

1) Pie Chart showing the total launches by a certain site/all sites. It displays1 relative proportions of multiple classes of data.

2) Scatter Graph showing the relationship between Outcome and Payload Mass(Kg) for different Launch Sites

Predictive Analysis (Classification)

1.Models:

- Logistic Regression
- Support Vector Machine
- Random Forest
- K-Nearest Neighbors

2.Pipeline:

- train_test_split(0.7/0.3)
- StandardScaler
- GridSearchCV

3.Validation: 5-fold CV

4.Top performer: Random Forest (≈ 85 % accuracy)

Results

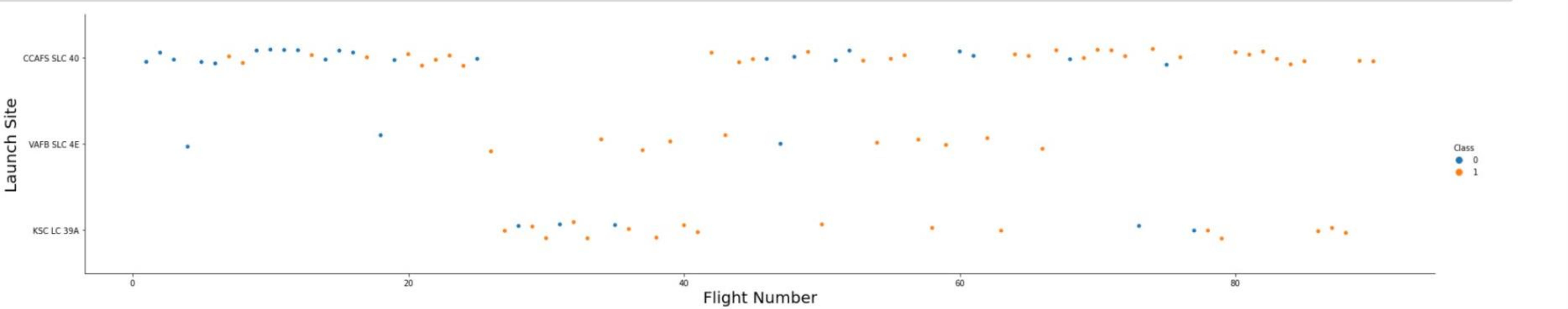
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

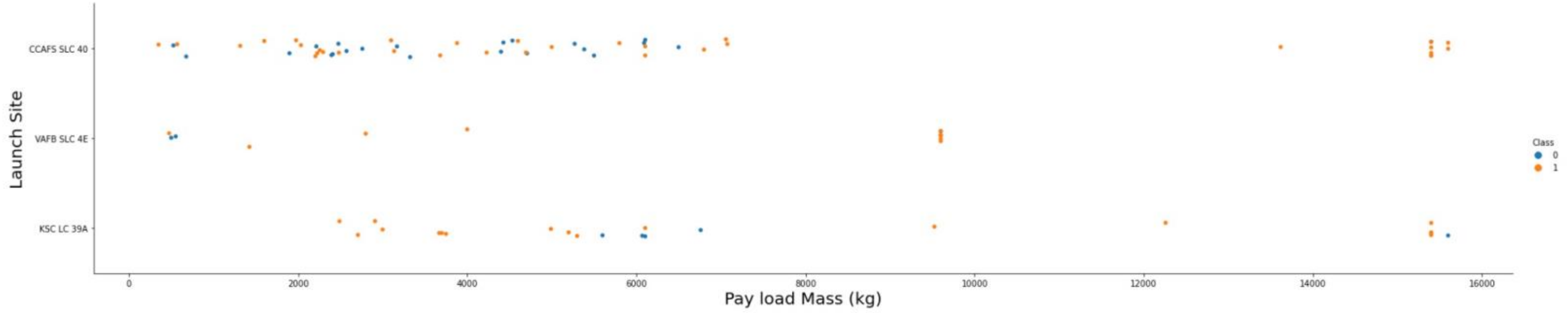
```
# Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hue to be the class value
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```



From the plot, we found that: larger the flight number at a launch site, greater is the success rate at a launch site.

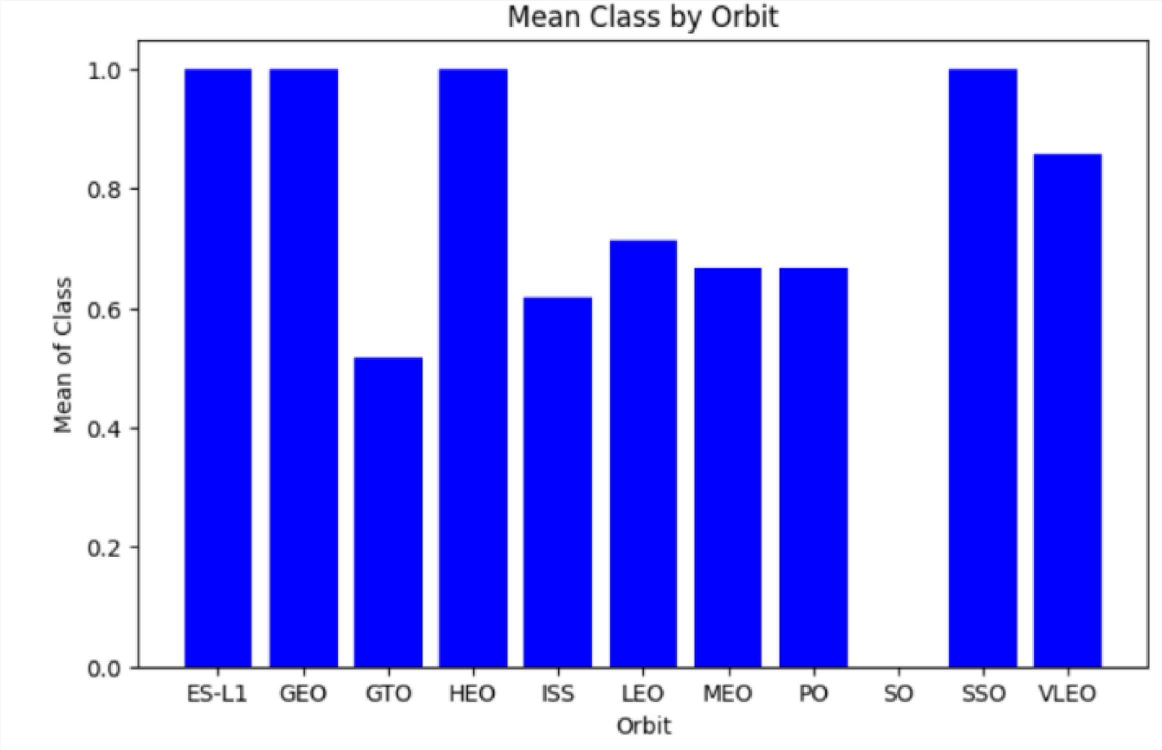
Payload vs. Launch Site

```python
# Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, and hue to be the class value
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("Pay load Mass (kg)",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```



Greater the payload mass for Launch Sites CCAFS SLC 40 and VAFB SLC 4E, higher the success rate for the Rocket.
There is not quite a clear pattern to be found using this visualization to make a decision if the Launch Site is dependant on Pay Load Mass for a success launch.

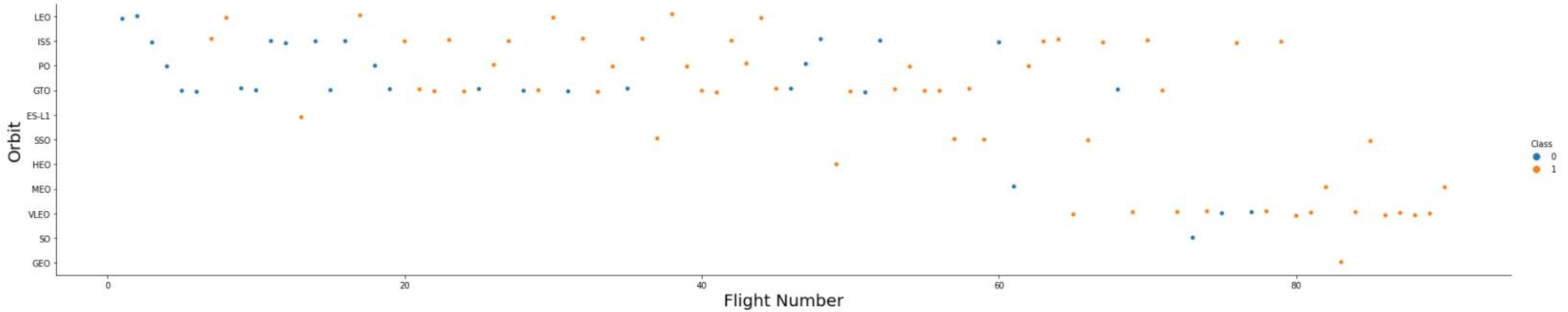
Success Rate vs. Orbit Type



- Orbit ES-L1, GEO, HEO, SSO and VLEO have the best Success Rates.

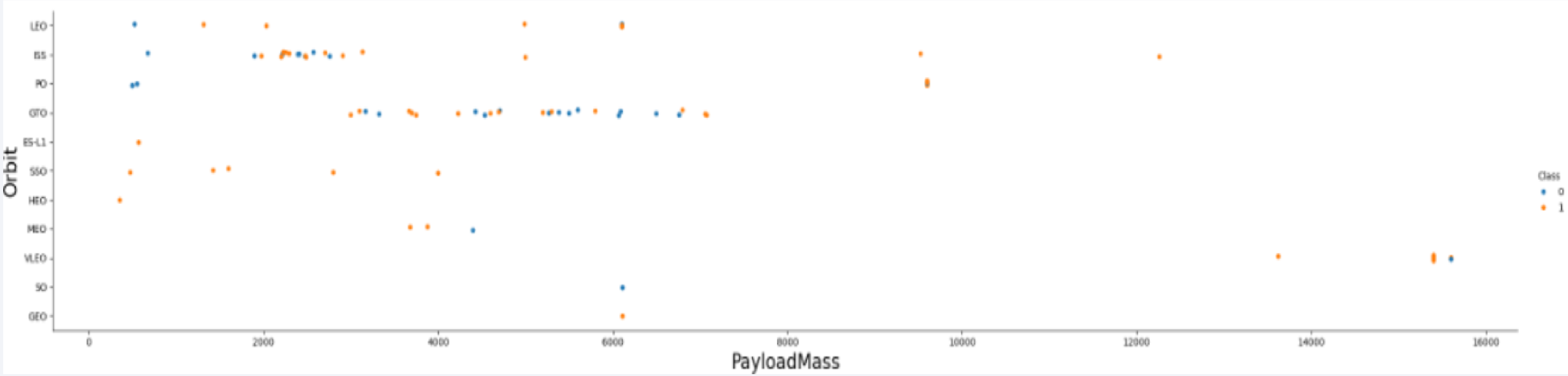
Flight Number vs. Orbit Type

```python
# Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be the class value
sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```



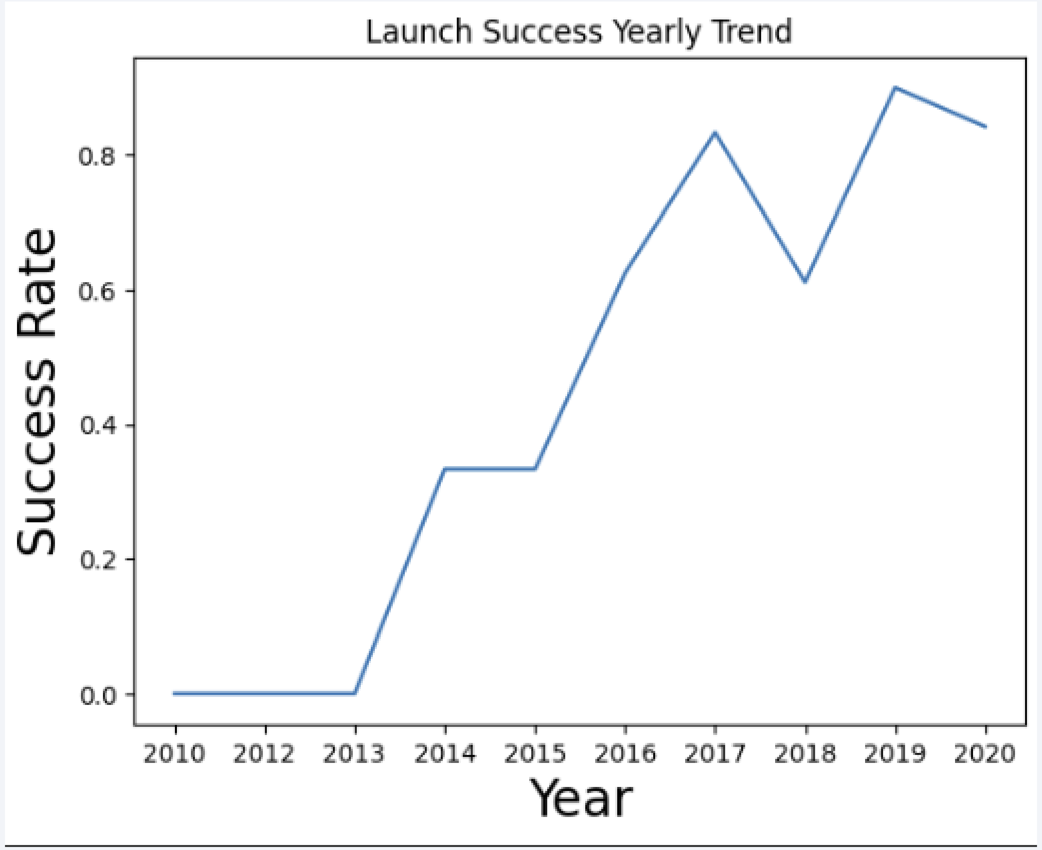
- In LEO orbit, the Success appears related to the flight number.
- SSO orbit had all flights as successful, but the data is less.
- There seems to be no relationship between flight number when in GTO and other orbits.

Payload vs. Orbit Type



- Heavy payloads have a positive influence on LEO, PO and ISS orbits.
- GTO orbit does not seem to have any relation with payload.

Launch Success Yearly Trend



All Launch Site Names

SQL query:

Keyword DISTINCT was used to show only unique launch sites from the SpaceX data.

- SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE

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

Launch Site Names Begin with 'CCA'

Query used to display 5 records where launch sites begin with `CCA`

- SELECT * from SPACEXTABLE 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

Total Payload Mass

Calculated the total payload carried for customer NASA:

- select sum("PAYLOAD_MASS__KG_") as "total_payload_mass" from SPACEXTABLE where "Customer" like "NASA (CRS)";

total_payload_mass
45596

Average Payload Mass by F9 v1.1

WHERE clause filters the dataset to only perform calculations on Booster_version F9 v1.1

- select AVG("PAYLOAD_MASS__KG_") as avg_payload_mass from SPACEXTABLE where "Booster_Version" like "F9 v1.1%";

avg_payload_mass
2534.666666666665

First Successful Ground Landing Date

‘Min’ function and ‘Where’ clause are used:

- select min("Date") from SPACEXTABLE where "Landing_Outcome" = "Success (ground pad)";

min("Date")
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Selecting only ‘Booster_Version ‘
- The WHERE clause filters the dataset to Landing_Outcome = Success (drone ship)
- The AND clause specifies additional filter conditions
- Payload_MASS_KG_>4000ANDPayload_MASS_KG_<6000
- select "Booster_Version" from SPACEXTABLE where "Landing_Outcome" = "Failure (drone ship)" AND "PAYLOAD_MASS__KG_"BETWEEN 1000 AND 5000;

Booster_Version
F9 v1.1 B1012
F9 v1.1 B1015
F9 FT B1024

Total Number of Successful and Failure Mission Outcomes

GROUP BY function used to group by 'Mission_Outcome' values and then values were counted.

- select "Mission_Outcome", COUNT(*) from SPACEXTABLE group by "Mission_Outcome";

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

Boosters Carried Maximum Payload

Subquery is used here:

- SELECT "Booster_Version" FROM SPACEXTABLE WHERE "Payload_Mass__kg_" = (SELECT MAX("Payload_Mass__kg_") FROM SPACEXTABLE);

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

2015 Launch Records

```
%sql SELECT substr(Date, 6,2) AS Month, "Landing_Outcome", "Booster_Version",
"Launch_Site" FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Failure (drone ship)' AND substr(Date,0,5)='2015';
```

```
 * sqlite:///my_data1.db
Done.
```

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

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

Function **COUNT** counts records in column, **WHERE** filters data and 'AND' is logical condition.

- SELECT Landing_Outcome, COUNT(Landing_Outcome) FROM SPACEXTABLE WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY COUNT(Landing_Outcome) DESC

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

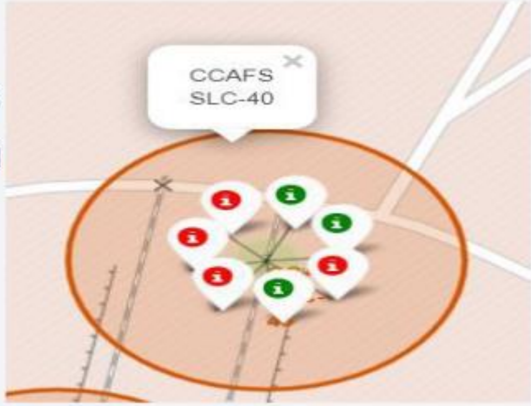
Section 3

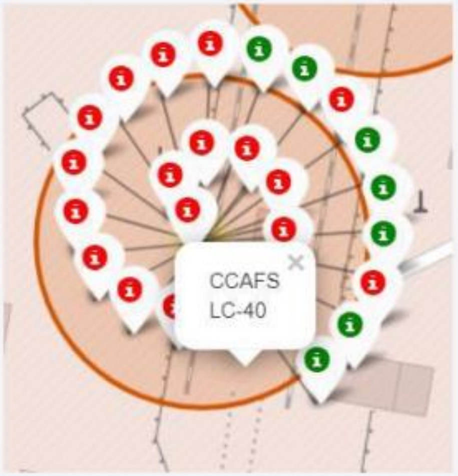
Launch Sites Proximities Analysis

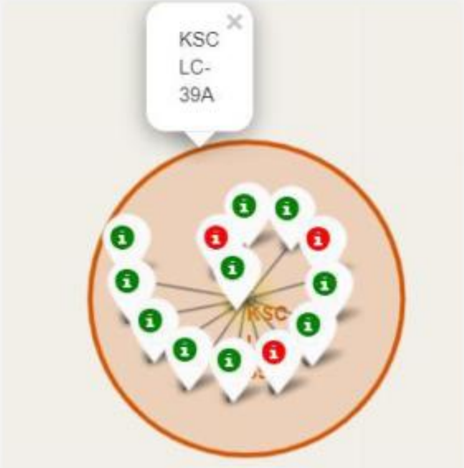
All launch sites in global map with markers

Markers with color labels for launch sites

Florida Launch Sites:



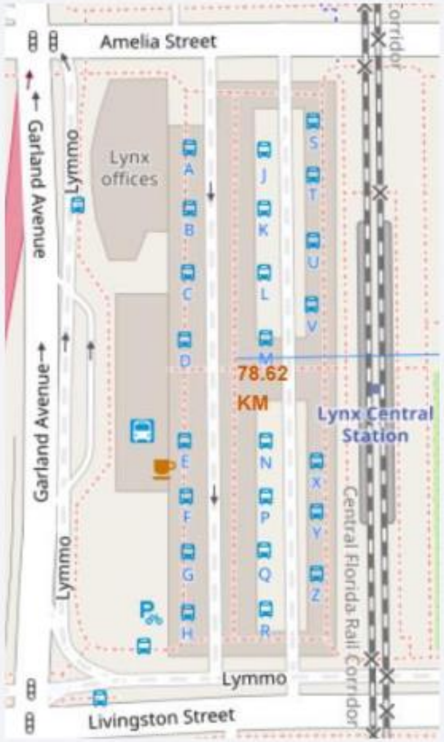




California Launch Sites:

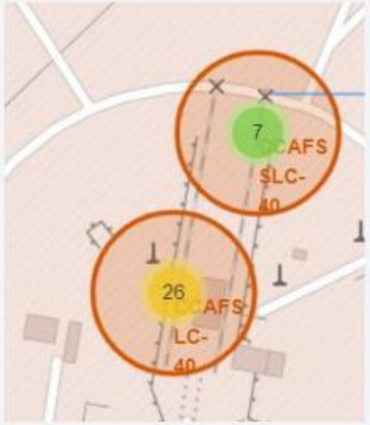
Green Markers: Successful Launches
Red Markers: Failed Launches

Distance of Launch sites to landmarks

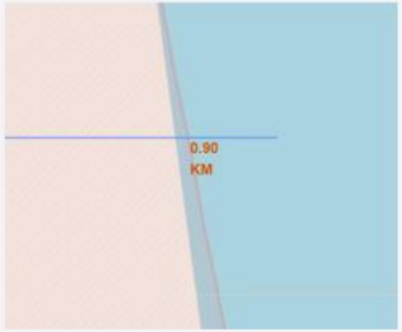


From Railway Station

Closest Highway



Coast



Coastline

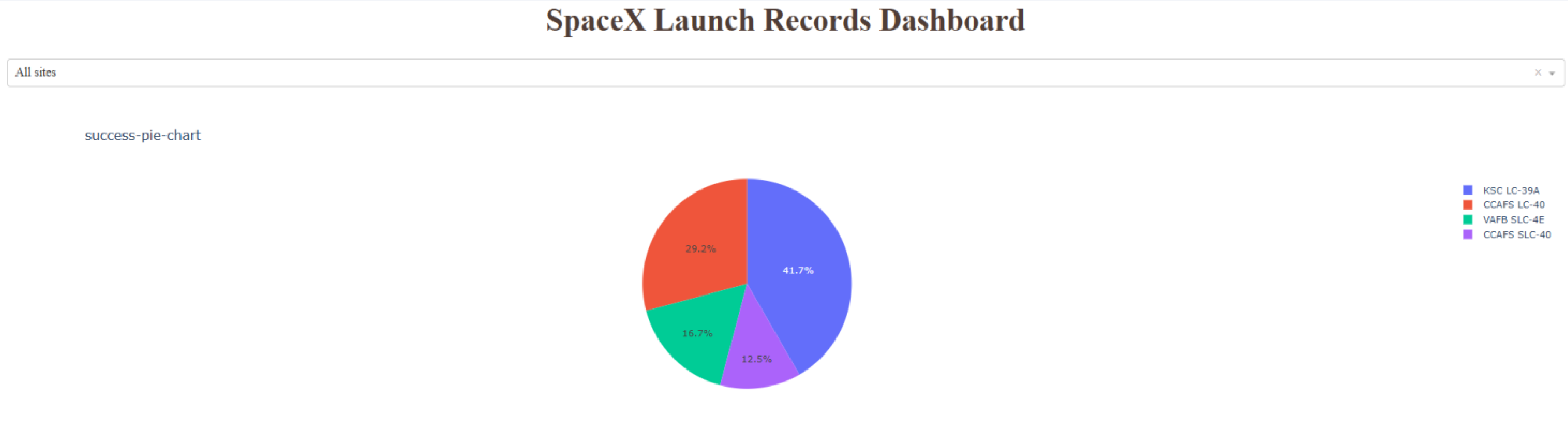
City

- Are launch sites in close proximity to railways? No
- Highway: No
- Coastline: Yes
- City: Away from cities

Section 4

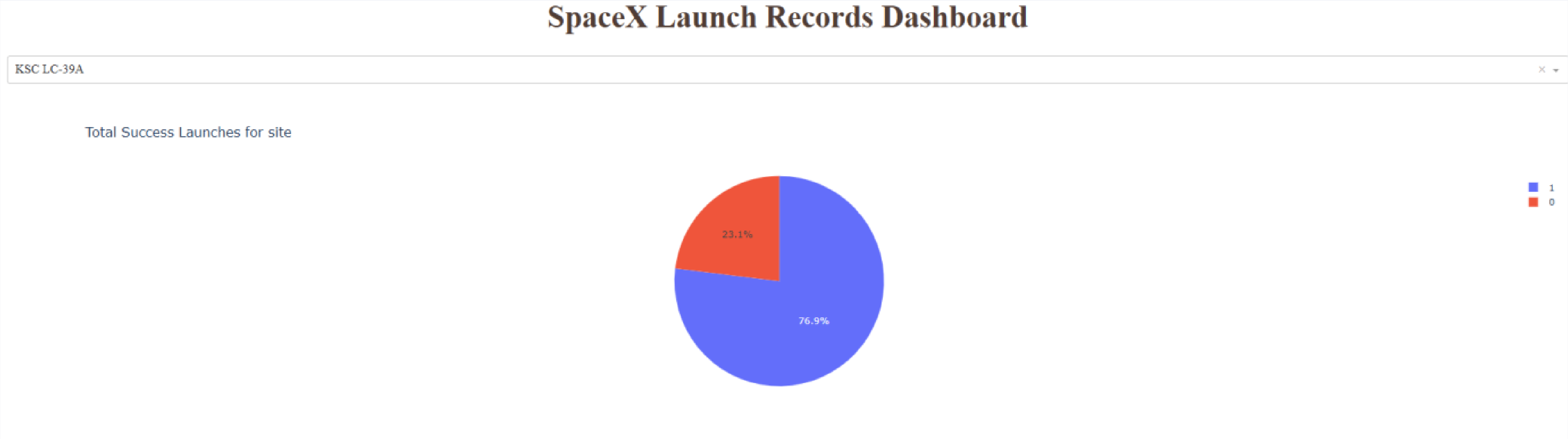
Build a Dashboard with Plotly Dash

Dashboard for total success launches by all sites



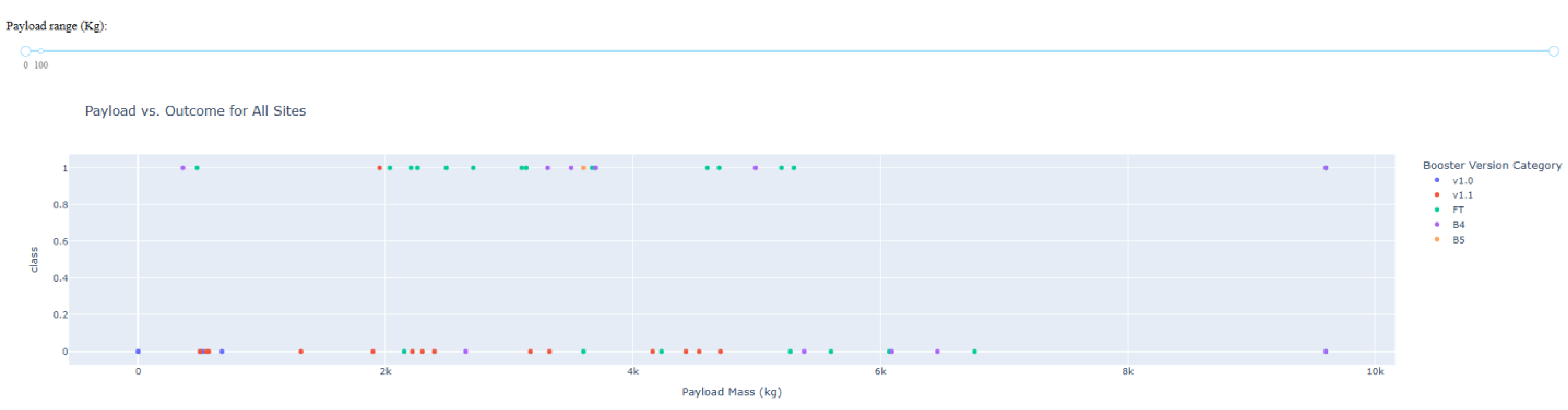
- KSC LC-39A had the most successful launches from all the sites

Pie chart for the launch site with highest launch success ratio



- KSC LC-39A has the highest success rate of 76.9%.

Payload vs. Launch Outcome scatter plot for all sites



- Success rate for V1.1 booster is less for all the payloads.
- Success rate of FT booster is good in general.
- Success rate for payloads more than 4000 kg is less .

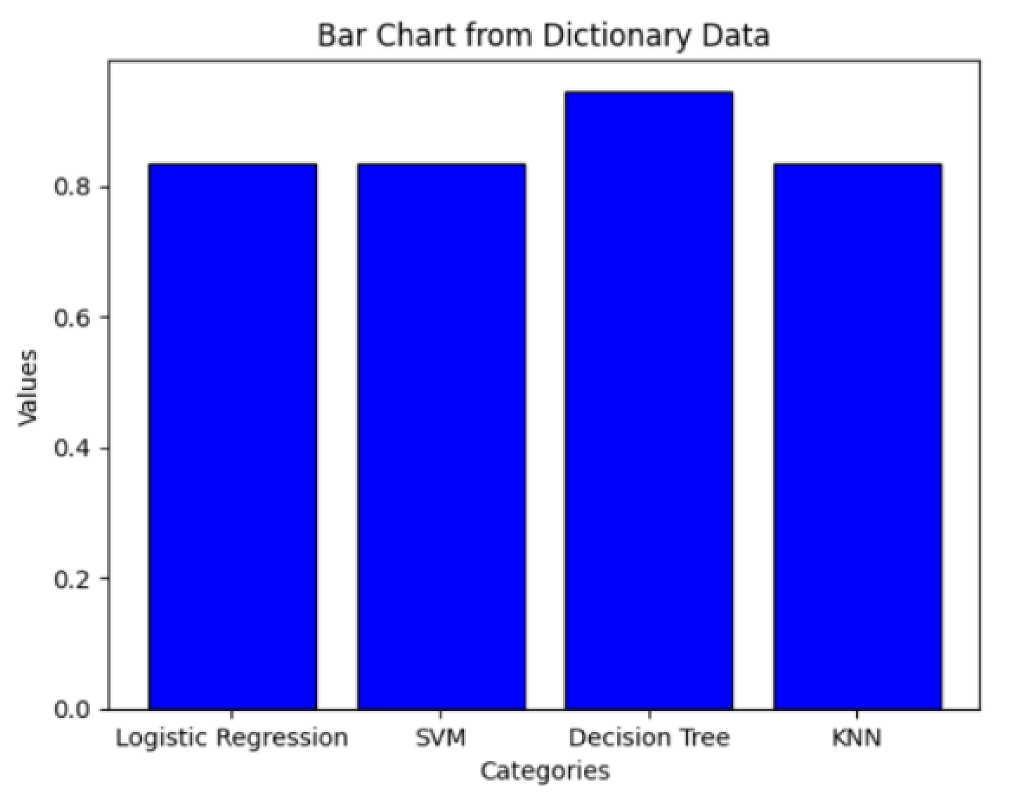
Section 5

Predictive Analysis (Classification)

Classification Accuracy

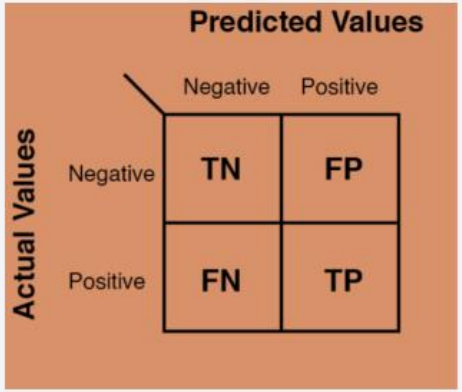
ML Model	Accuracy
Logistic regression	0.83333
SVM	0.83333
Decision Tree	0.94444
KNN	0.83333

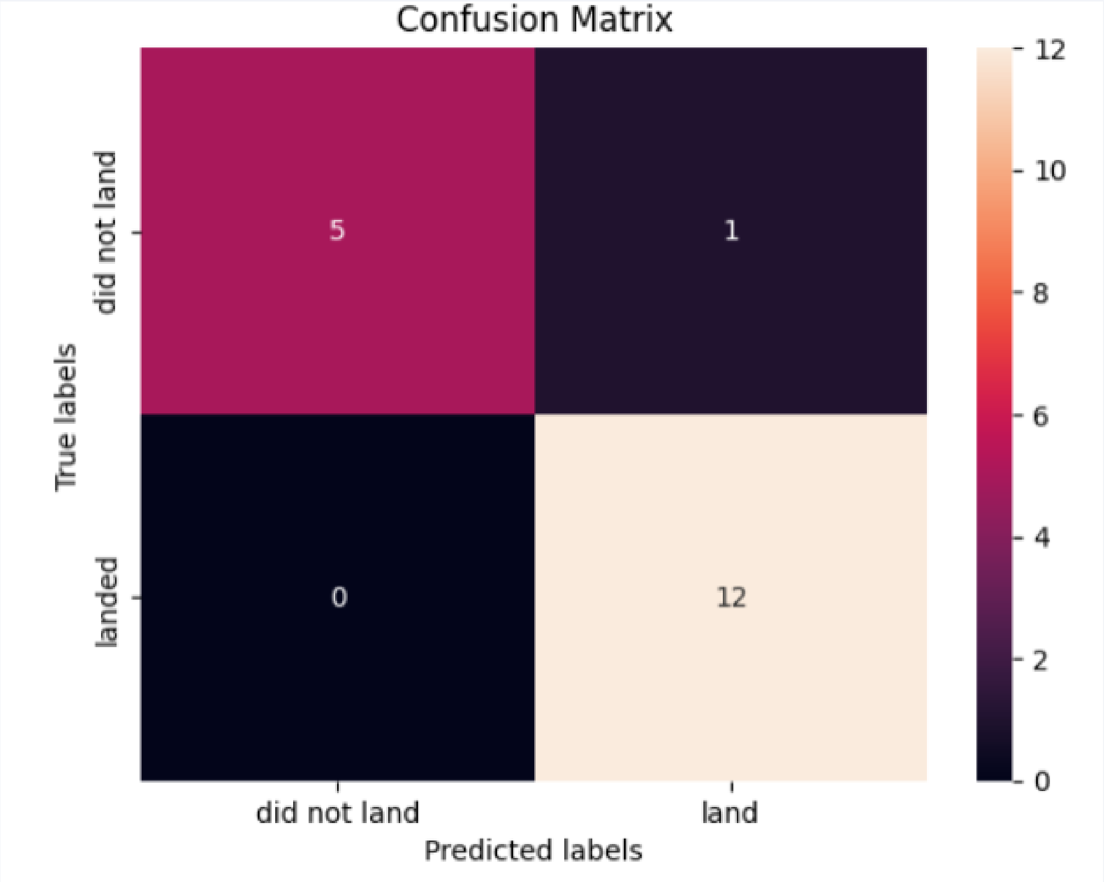
- Decision Tree model have highest accuracy with test data.



Confusion Matrix

- False Positive prediction is 1.
- Other than this result, decision tree's predictions are accurate.
- Hence, we can say that the decision tree model performs well.





Conclusions

- As the time passed in months and years, success rate of SpaceX launches got better.
- Success rate for payloads more than 4000 kg is less.
- Success rate for V1.1 booster is less for all the payloads, whereas success rate of FT booster is high.
- KSC LC-39A has the highest success rate of 76.9%.
- Launch sites are located near coastlines.
- Heavy payloads have a positive influence on LEO, PO and ISS orbits
- Orbit ES-L1, GEO, HEO, SSO and VLEO have the best success Rates.
- Larger the flight number at a launch site, greater is the success rate at a launch site.
- Built a Plotly interactive dashboard to show pie chart and scatter plots of launch success rate for different launch sites.
- Decision tree ML algorithm was best prediction algorithm for this dataset.

Thank you!