Applied Data Science Capstone

Outline

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

Executive Summary

- Methodologies Summary
  - Data Collection : link to GitHub
  - Data Wrangling : link to GitHub
  - EDA Data Visualization : link to GitHub
  - EDA with SQL : link to GitHub
  - Predictive Analysis : link to GitHub
  - Interactive Map with Folium : link to GitHub
  - Interactive Dashboard Using Plotly

Results Summary

- Exploratory Data Analysis
- Predictive Analysis Results

Introduction

- Background and Problem

SpaceX is an aerospace manufacturer and space transportation services and communications corporation, much like Tesla, both founded by Elon Musk. Despite being less than 20 years old, SpaceX has managed to reduce launch costs by more than 50% compared to other companies, with projections suggesting a 99% reduction once the Starship project is completed. This reduction is largely attributed to SpaceX's development of technology to land the first stage booster, which constitutes 70% of the rocket's cost. By safely landing and reusing the booster, SpaceX significantly cuts down on launch expenses. Reusing boosters reduces costs by 50% compared to using new boosters, solidifying SpaceX's dominance in the market. In this capstone project, we will analyze data extracted from Wikipedia through web scraping and the SpaceX API to gain insights and predict safe booster landings onto drone ships.

Methodology

- Data Collection : API & Web Scraping
- Data Wrangling : Extracting Load & Transform
- Cleaning Data
- EDA with Visualization and SQL
- Interactive Dashboard using Folium and Plotly Dash
- Predictive Analysis using Machine Learning Models.

Methodology

- Data Collection
  - REST API : Using the rest API we extract the data in form of JSON and transform it to a data frame using inbuilt python pandas method normalize.
  - Web Scraping : Web scraping SpaceX launches from Wikipedia and converting it into a data frame.

Rest API

- Make Request

```
json_data=requests.get(static_json_url).json()
```

- Normalize to df

```
# Use json_normalize meethod to convert the json result into a dataframe
data=pd.json_normalize(json_data)
```

- Create a dictionary for dataframe

```
launch_dict = {'FlightNumber': list(data['flight_number']),
'Date': list(data['date']),
'BoosterVersion':BoosterVersion,
'PayloadMass':PayloadMass,
'Orbit':Orbit,
'LaunchSite':LaunchSite,
'Outcome':Outcome,
'Flights':Flights,
'GridFins':GridFins,
'Reused':Reused,
'Legs':Legs,
'LandingPad':LandingPad,
'Block':Block,
'ReusedCount':ReusedCount,
'Serial':Serial,
'Longitude': Longitude,
'Latitude': Latitude}
```

- Save to CSV

```
data_falcon9.to_csv('dataset_part_1.csv',index=False)
```

- Filter

```
# Hint data['BoosterVersion']!='Falcon 1'
data_falcon9=df_launch[df_launch['BoosterVersion']!='Falcon 1']
```

Data Wrangling

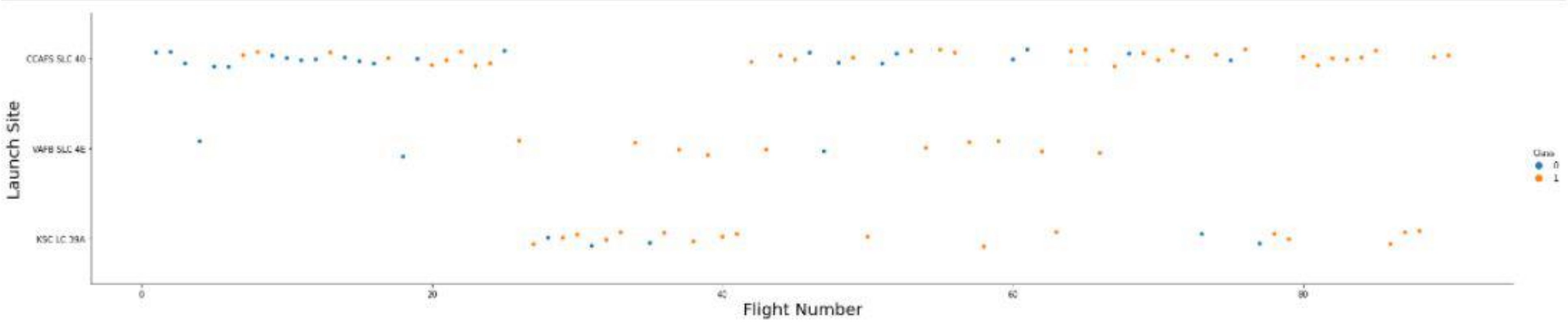
In the dataset, there are various instances where the booster did not land successfully. Sometimes, a landing was attempted but failed due to an accident. For instance, "True Ocean" indicates a mission outcome where the landing was successful in a specific region of the ocean, whereas "False Ocean" indicates an unsuccessful landing in a specific region of the ocean. Similarly, "True RTLS" denotes a mission outcome where the landing was successful on a ground pad, whereas "False RTLS" indicates an unsuccessful landing on a ground pad. "True ASDS" signifies a mission outcome where the landing was successful on a drone ship, while "False ASDS" indicates an unsuccessful landing on a drone ship.

Exploratory Data Analysis with Visualization

Through Exploratory Data Analysis (EDA) using data from the API and Wikipedia, we will derive insights on the following:

- Visualization of launch frequency from each launch site based on flight number.
- Analysis of payloads launched from different sites.
- Comparison of success rates across different orbit types.
- Examination of orbit types associated with each flight number.
- Exploration of payload types and their corresponding orbits.
- Examination of the trend in success rates over the years.

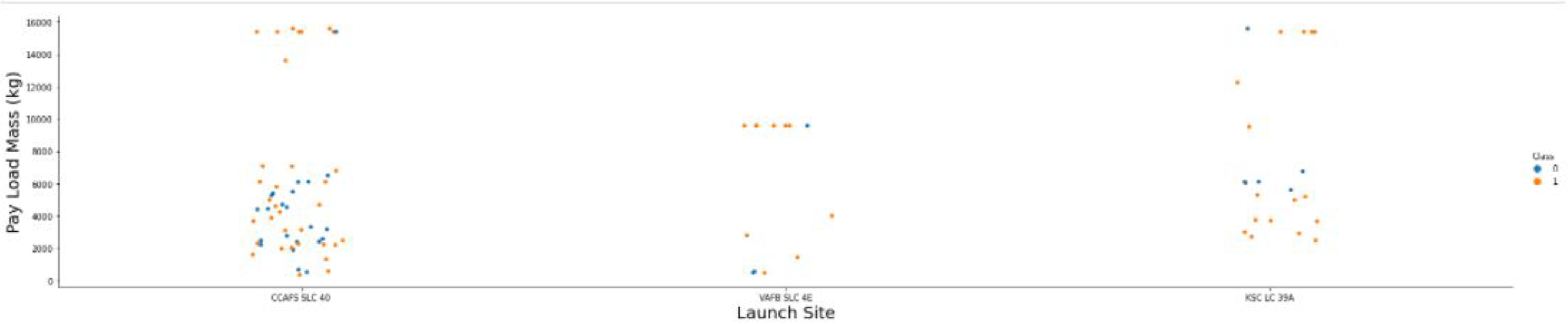
Flight Number & Launch Sites



Based on the visualization, we can conclude the following:

- Earlier flights were predominantly launched from the CCAFS-SLC-40 site, followed by KSC-LC-39A.
- The majority of launches are conducted from the CCAFS-SLC-40 site.
- There are fewer launches from the VAFB SLC 4E site.

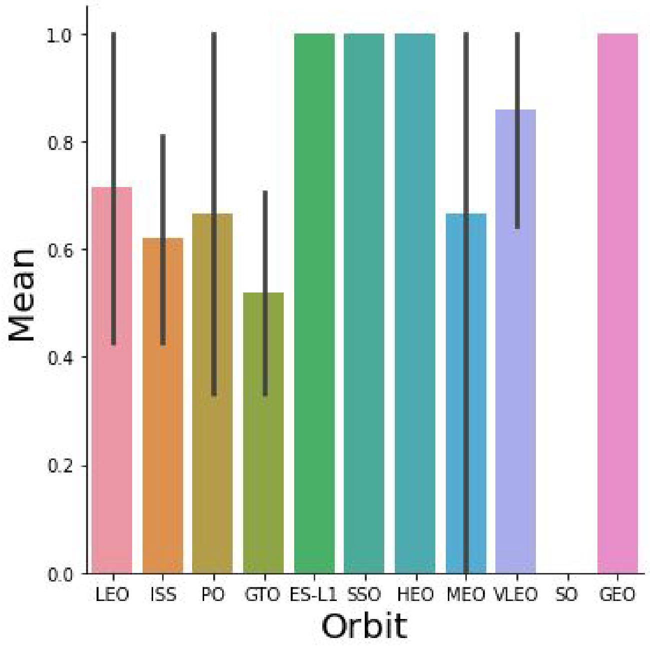
Payload & Launch Sites



Based on the visualization, we can conclude the following:

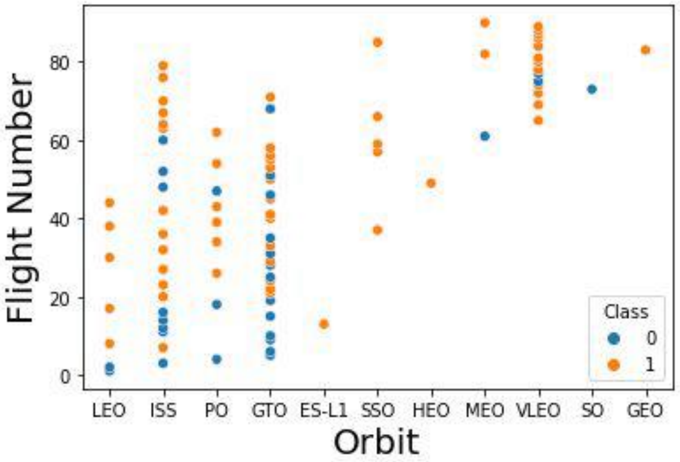
- VAFB SLC 4E exhibits lower payload launches compared to other sites.
- CCAFS SLC 40 has a mix of both higher payload launches and lower payload launches, indicating versatility in mission types conducted from this site.

Orbit Success



Based on the visualization, we can conclude that GEO (Geostationary Orbit), HEO (Highly Elliptical Orbit), ES-L1 (Earth-Sun L1), and SSO (Sun-Synchronous Orbit) have a high success rate compared to other orbit types.

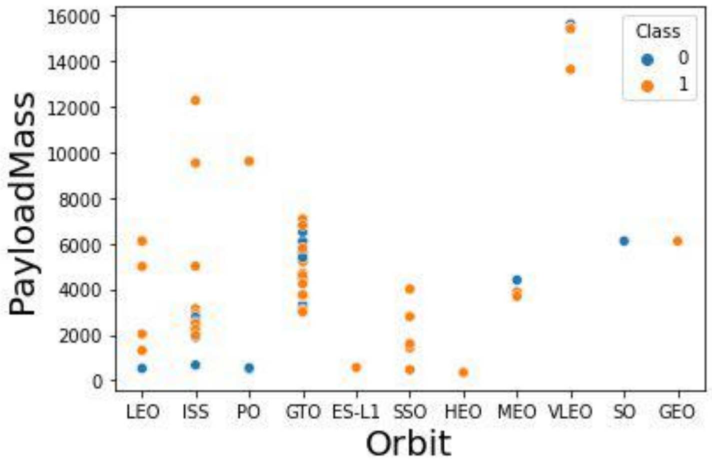
Flight Number & Orbit



Based on the visualization, we can conclude the following insights:

- Most flights are to ISS (International Space Station), Polar Orbit (PO), Geostationary Transfer Orbit (GTO), and Very Low Earth Orbit (VLEO).
- The majority of failures occur for missions to ISS and GTO.
- Sun-Synchronous Orbit (SSO) and Very Low Earth Orbit (VLEO) exhibit a high success rate compared to other orbit types.

Payload & Orbit

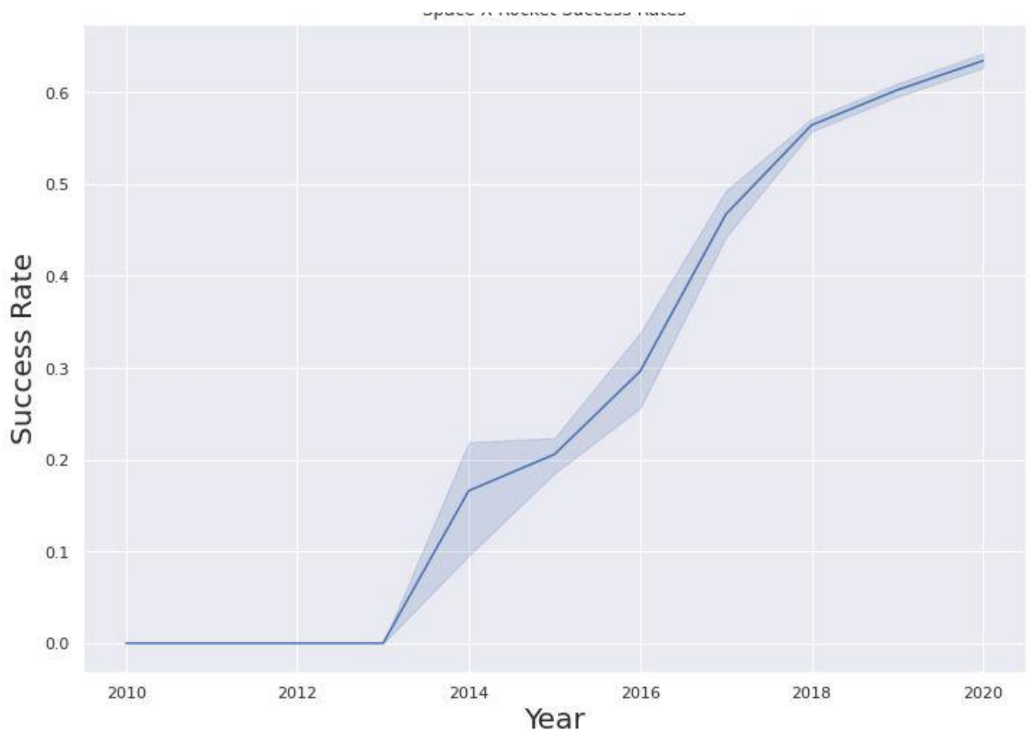


Based on the visualization, we can draw the following conclusions:

- Higher payloads are typically associated with launches to Very Low Earth Orbit (VLEO).
- Launches to Highly Elliptical Orbit (HEO), International Space Station (ISS), Polar Orbit (PO), and Earth-Sun L1 (ES-L1) tend to have lower payload sizes.
- Geostationary Transfer Orbit (GTO) generally sees launches with an average payload size.

Success Rate Trend

The success rate of launches increases over time due to the insights gained from analyzing past failures and successes in the collected data.



Exploratory Data Analysis with SQL

- Exploratory Data Analysis on the follow criteria:
- Unique Sites
- Max Payload
- Average Payload
- Day when First Success Landing
- Success and Failures count
- Boosters With Max Payload

Exploratory Data Analysis with SQL

For the categories above we find that :

- Sites that SpaceX operates in are:
  - CCAFS LC-40,CCAFS SLC-40,KSC LC-39A,VAFB SLC-4E
- Max Payload: 48213
- Average Payload for all Launches: 2928 Kgs
- First Success Landing was Made on: 06/05/2016
- Booster Version that carry over 4000 kg and 6000 Kg :
  - F9 FT B1020, F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2

Interactive Map with Folium

Visualization of the launches for every site and every launch in a Interactive Map

- Visualization of:
  - Launch Sites
  - Visualize the launches on the map base on Fail or Success

Predictive Analysis

- Through Models, tuned for best performance we got the insights on the probability if a launch being success or a failure .
- Models used include:
  - KNeighboursClassfier
  - Decision Tree
  - Logistic Regression
  - Support Vector Machine

Best Model

After Analyzing all the Models, the KNN method was the best Model with accuracy of 77% and best Score of 87%

```
parameters = {'n_neighbors': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
              'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute'],
              'p': [1,2]}

KNN = KNeighborsClassifier()
gscv=GridSearchCV(KNN,parameters,scoring="accuracy",cv=10)
KNN_cv=gscv.fit(X_train,y_train)
```

```
print("Accuracy",KNN_cv.score(X_test,y_test))
```

```
Accuracy 0.7777777777777778
```

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

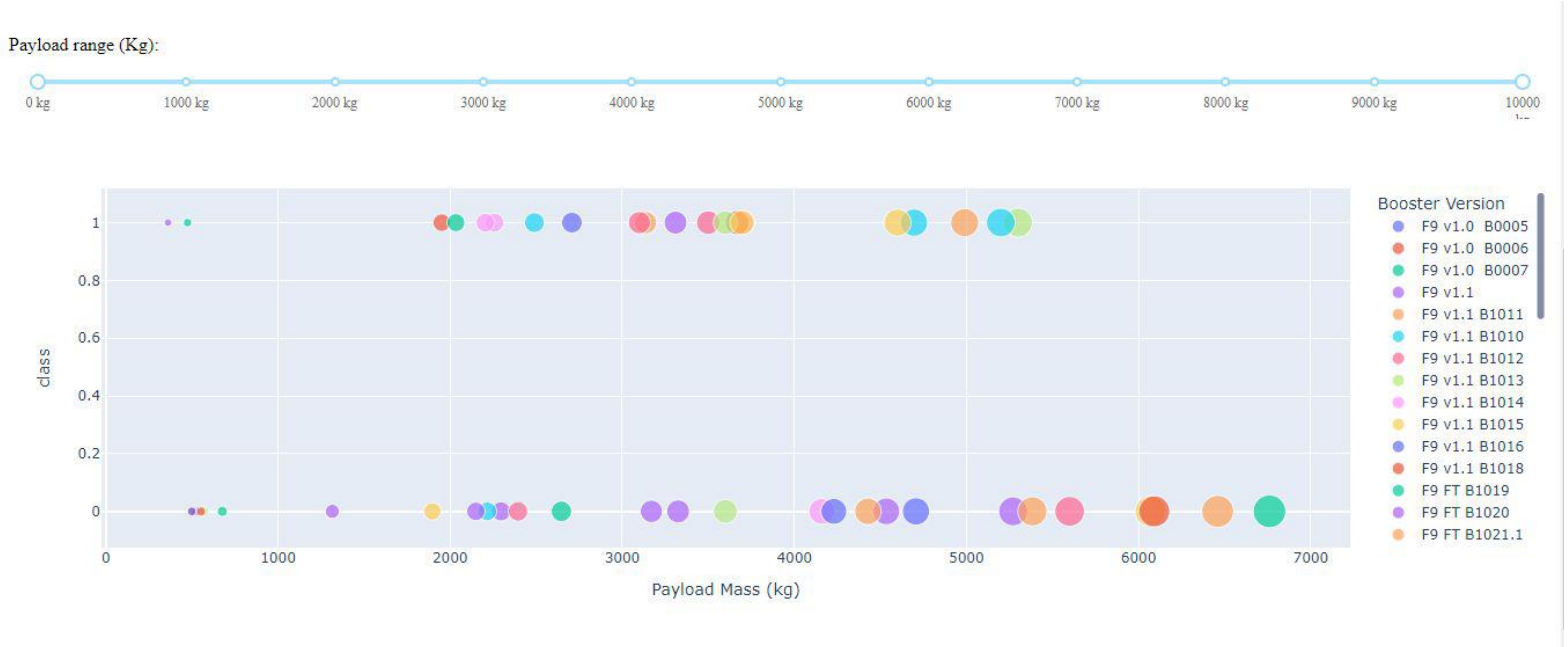
```
tuned hpyerparameters :(best parameters)  {'algorithm': 'auto', 'n_neighbors': 4, 'p': 1}
accuracy : 0.8767857142857143
```

Interactive with Dash

Total Success Launches By all sites

The observation indicates that Kennedy Space Center (KSC) has more launches compared to other sites. By utilizing the drop-down menu on the dashboard, it is possible to view launches specifically from a single site, allowing for focused analysis and visualization of site-specific launch data.

Success Launches for site VAFB SLC-4E



By utilizing the range slider, we can analyze and visualize the success and failure of launches for each booster version and the payload they were carrying. This interactive tool enables us to explore the relationship between booster versions, payload types, and mission outcomes, providing valuable insights into performance across different parameters.