Applied Data Science Capstone

Executive Summary

- This project investigates SpaceX's Falcon 9 launch data to determine which launch sites and configurations yield successful landings. By applying data wrangling, EDA, visualization techniques, and machine learning models, we aim to identify patterns and improve the prediction of mission outcomes. Interactive dashboards and maps enhance accessibility and interpretation of our results.

Introduction

- SpaceX aims to reduce space transportation costs and enable the colonization of Mars. This project focuses on Falcon 9 launches and examines features like launch site, orbit, payload mass, and landing outcomes. The objective is to derive insights from historical data and build models to predict mission success.

Data Collection & Wrangling

- **Data Source:** SpaceX Falcon 9 dataset (CSV) provided by IBM on Coursera.
- **Tools Used:** Python, Pandas, NumPy.
- **Key Steps:**
  - Read datasets using pandas.read_csv()

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857
5	6	2014-01-06	Falcon 9	3325.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1005	-80.577366	28.561857
6	7	2014-04-18	Falcon 9	2296.000000	ISS	CCAFS SLC 40	True Ocean	1	False	False	True	NaN	1.0	0	B1006	-80.577366	28.561857
7	8	2014-07-14	Falcon 9	1316.000000	LEO	CCAFS SLC 40	True Ocean	1	False	False	True	NaN	1.0	0	B1007	-80.577366	28.561857
8	9	2014-08-05	Falcon 9	4535.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1008	-80.577366	28.561857
9	10	2014-09-07	Falcon 9	4428.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1011	-80.577366	28.561857

Data Collection & Wrangling

- Handled missing values, dropped irrelevant columns (id, Unnamed: 0)

```
FlightNumber      0.000000
Date              0.000000
BoosterVersion    0.000000
PayloadMass       0.000000
Orbit             0.000000
LaunchSite        0.000000
Outcome           0.000000
Flights           0.000000
GridFins          0.000000
Reused            0.000000
Legs              0.000000
LandingPad       28.888889
Block             0.000000
ReusedCount       0.000000
Serial            0.000000
Longitude         0.000000
Latitude          0.000000
dtype: float64
```

- Created new feature Landing_Class based on mission outcome.

```
LaunchSite
CCAFS SLC 40    55
KSC LC 39A      22
VAFB SLC 4E     13
Name: count, dtype: int64
```

EDA Methodology

- Univariate analysis to understand distributions
- Value counts for LaunchSite, Orbit, and Outcome
- Boxplots to detect price outliers
- Scatter plots with seaborn.regplot for correlation checks
- Groupby analysis for orbit and mission outcome trends

EDA Methodology

- Univariate analysis to understand distributions

```
FlightNumber        int64
Date               object
BoosterVersion     object
PayloadMass       float64
Orbit              object
LaunchSite         object
Outcome            object
Flights             int64
GridFins             bool
Reused               bool
Legs                 bool
LandingPad         object
Block             float64
ReusedCount         int64
Serial             object
Longitude         float64
Latitude          float64
dtype: object
```

- Value counts for LaunchSite, Orbit, and Outcome

```
Orbit
GTO     27
ISS     21
VLEO    14
PO       9
LEO      7
SSO      5
MEO      3
ES-L1    1
HEO      1
SO       1
GEO      1
Name: count, dtype: int64
```

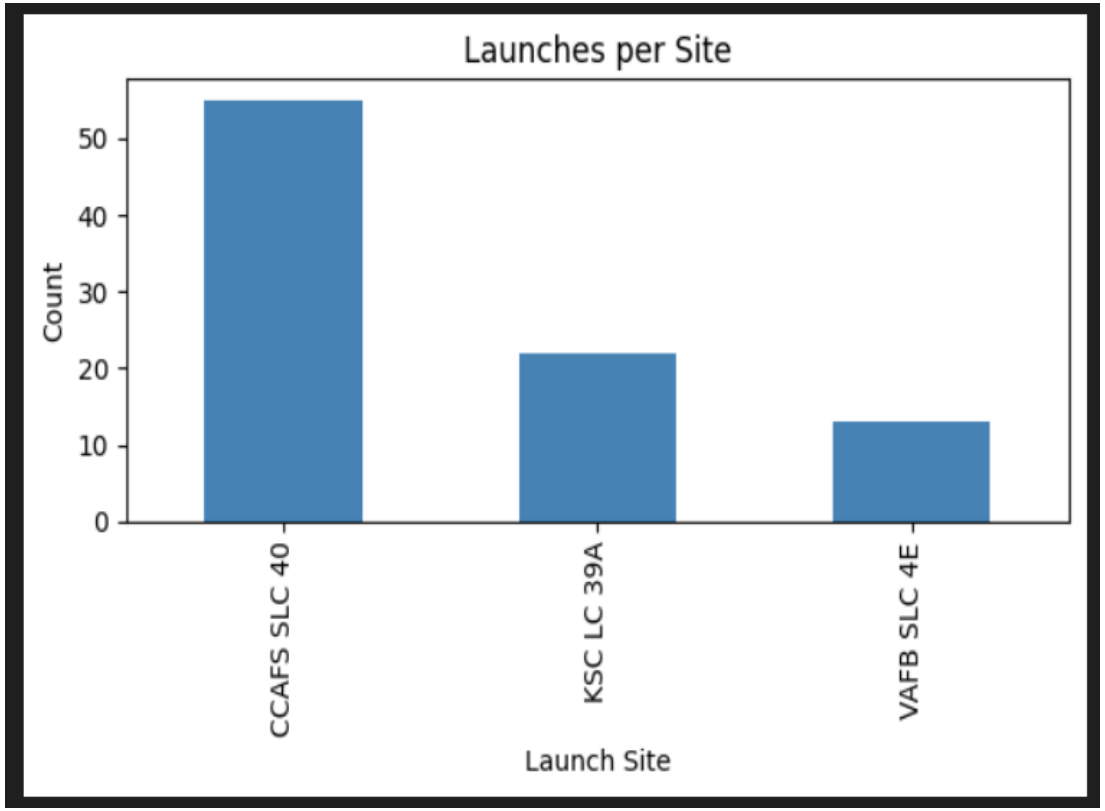
```
Outcome
True ASDS      41
None None      19
True RTLS      14
False ASDS      6
True Ocean      5
False Ocean     2
None ASDS       2
False RTLS      1
Name: count, dtype: int64
```

```
0 True ASDS
1 None None
2 True RTLS
3 False ASDS
4 True Ocean
5 False Ocean
6 None ASDS
7 False RTLS
```

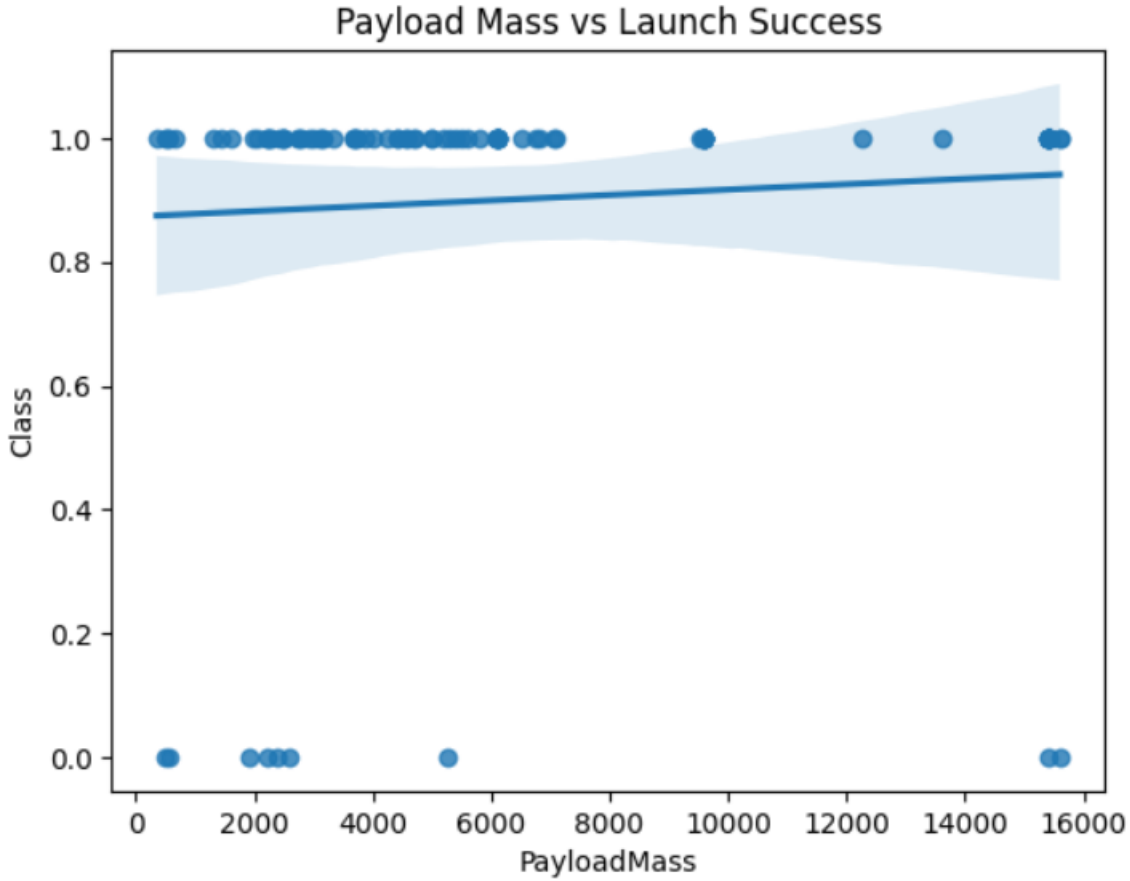
```
{'False ASDS', 'False Ocean', 'False RTLS', 'None ASDS', 'None None'}
```

EDA Methodology

- Boxplots to detect price outliers

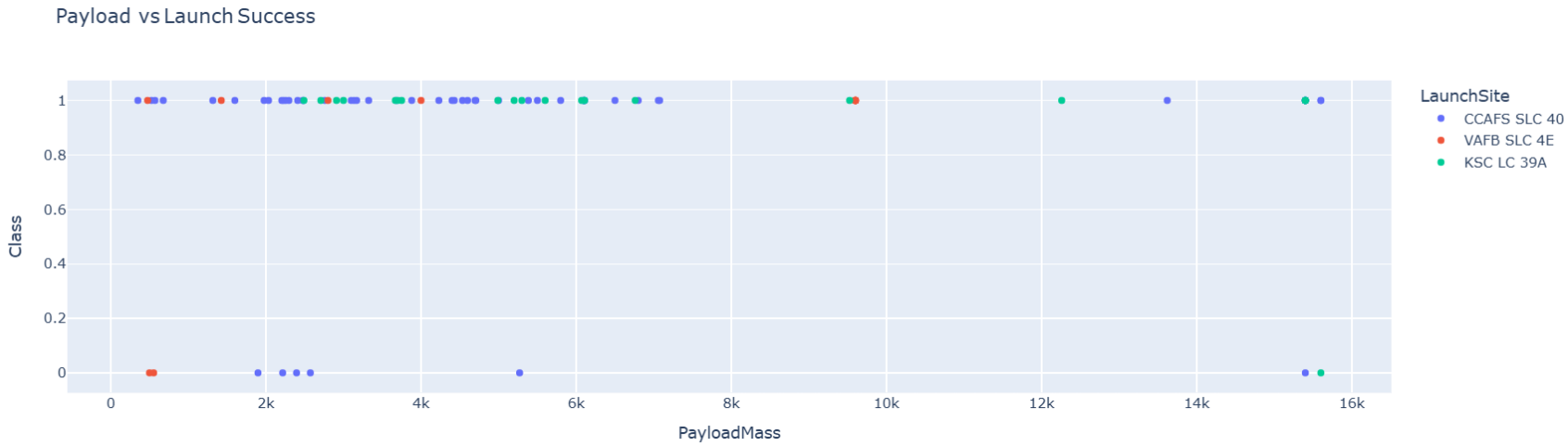


- Scatter plots with seaborn.regplot for correlation checks



Interactive Visual Analytics Methodology

- Used Folium to generate interactive maps
- Used Plotly for interactive charts
- Dashboards created using Plotly Dash for real-time filtering and analysis



Predictive Analysis Methodology

- Regression and classification models used
- Linear Regression: predict PayloadMass from FlightNumber

```
Linear Regression R² on test set: -0.5600762251690348
```

- Ridge Regression with polynomial transformation

```
Ridge (Poly deg 2) R² on test set: -0.8187567869903685
```

- Logistic Regression to classify mission outcome (Landing Success)

```
Logistic Regression accuracy: 0.8888888888888888
Confusion matrix:
 [[ 4  2]
 [ 0 12]]
```

EDA Visualization Results (1/2)

- LaunchSite Frequency:
  - CCAFS LC-40: Most frequent
  - KSC LC-39A & VAFB SLC-4E follow
- Orbit Types:
  - LEO most common, followed by GTO and ISS
- Mission Outcomes:
  - High success with True ASDS, True RTLS
  - None and False represent failures

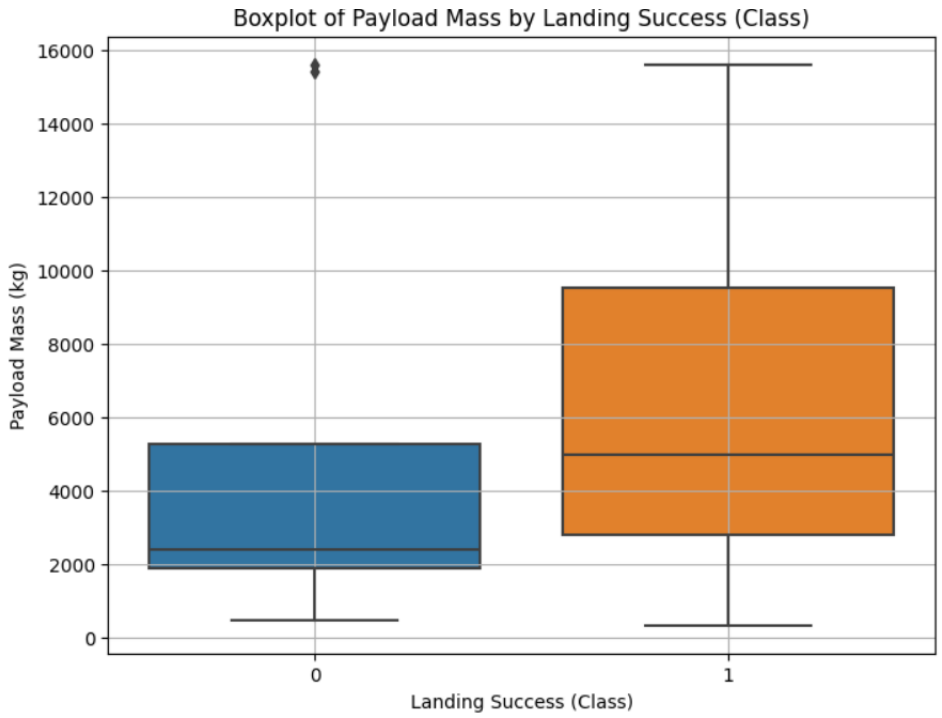
```
0 True ASDS
1 None None
2 True RTLS
3 False ASDS
4 True Ocean
5 False Ocean
6 None ASDS
7 False RTLS
```

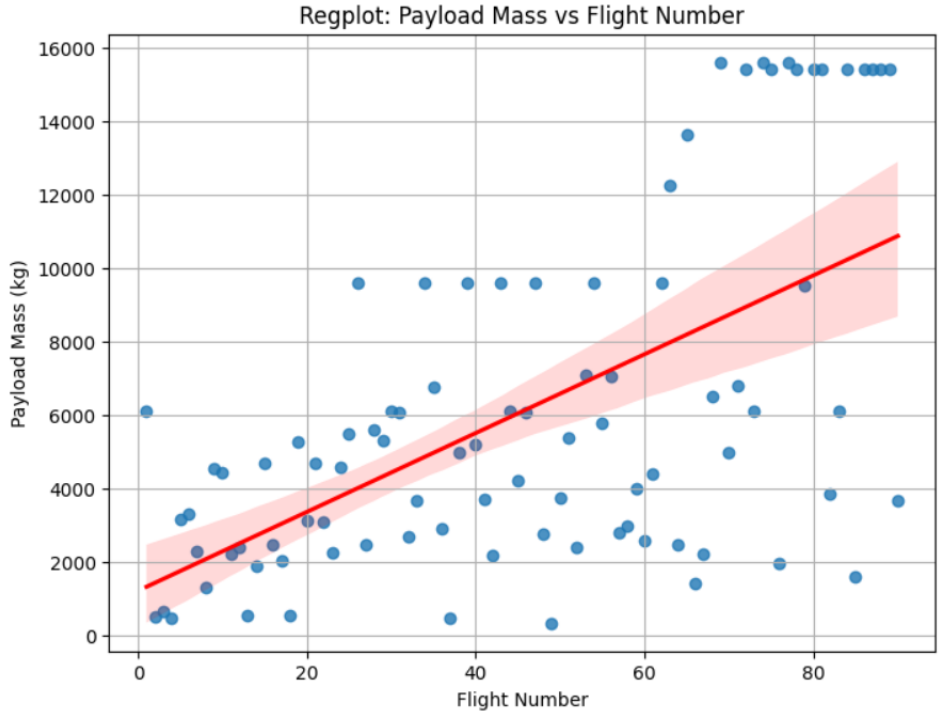
EDA Visualization Results (2/2)

Boxplot: More outliers in **PayloadMass** for **failed landings** (Class = 0) than for successful ones (Class = 1)

Regplot: PayloadMass is positively correlated with **FlightNumber**

- **Features like BoosterVersion and FlightNumber show consistent trends**, but **PayloadMass** plays a significant role in launch outcomes





SQL Analysis Results

- Total launches to GEO: 27
- Total successful landings on drone ships (True ASDS): 41
- Launch counts by site, orbit distribution, and outcome analysis using pandasql and SQL queries on sqlite3

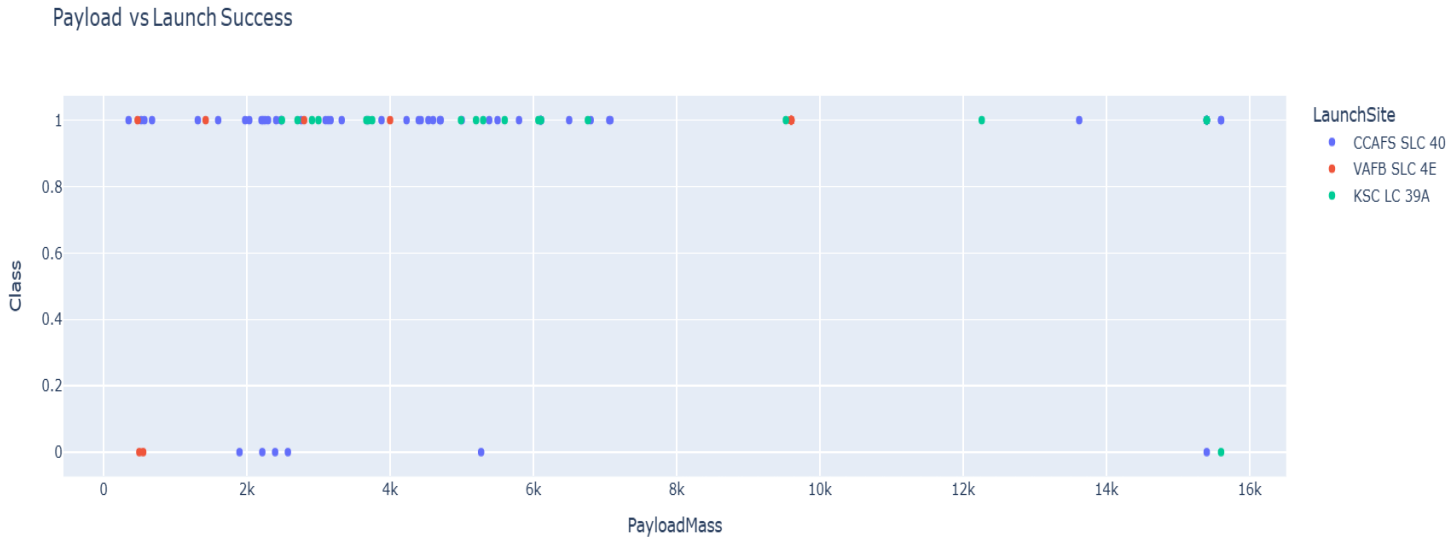
```
Total launches to GEO: 27
Total successful landings on drone ships (True ASDS): 41
```

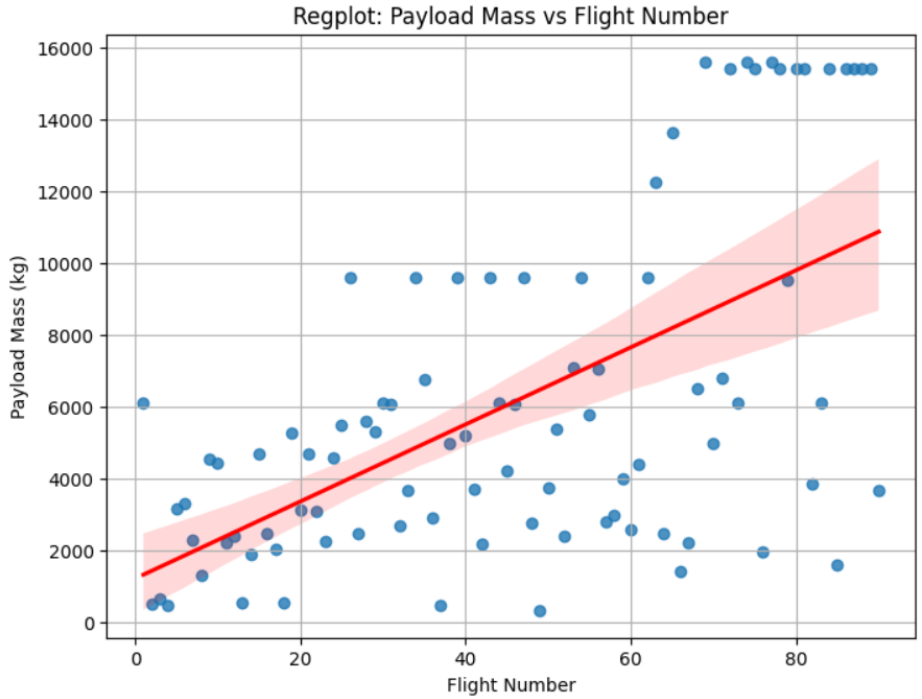
Folium Interactive Map Results

- Visualized all SpaceX launch sites on an interactive world map
- Clickable markers display location and site info
- Shows geographical trends in launch site usage

Plotly Dash Dashboard Results

- Interactive dropdown to select launch site
- Pie chart for successful vs failed missions
- Scatter plot of payload mass vs mission outcome
- Easy-to-use UI built with callbacks in Dash





Predictive Analysis Results

- Regression and classification models used
- Linear Regression: predict PayloadMass from FlightNumber

```
Linear Regression R² on test set: -0.5600762251690348
```

- Ridge Regression with polynomial transformation

```
Ridge (Poly deg 2) R² on test set: -0.8187567869903685
```

- Logistic Regression to classify mission outcome (Landing Success)

```
Logistic Regression accuracy: 0.8888888888888888
Confusion matrix:
 [[ 4  2]
 [ 0 12]]
```

Conclusion

- Launch site and orbit type significantly affect mission success.
- Strong predictors of successful landing: payload mass, orbit, and site.
- Interactive visualizations and models enhance interpretability.
- Predictive models offer valuable insights for future SpaceX missions.

Creativity & Insights

- Added real-time interactivity using Dash and Folium
- Inferred correlation between site configuration and success rate
- Suggested deployment of dashboard for operations team at SpaceX