Capstone Project: Data Science Journey

Arwa Abbas, 14-Sep-2025, IBM Data Science Professional Certificate: Data Science Capstone Course

Executive Summary

•**Objective:** Apply full data-science workflow to a real dataset

•**Tools:** Python, Pandas, SQL, Folium, Plotly Dash, scikit-learn

•**Key Result:** To built an interactive dashboard & classification model with solid accuracy

•**Outcomes:** To support data-driven decisions

Introduction

- Overview of SpaceX launches (2010–2020)
- Columns in the dataset (Launch Site, Payload Mass, Orbit, Class, Landing Outcome)
- Motivation: Improve understanding of success/failure trends and predict outcomes.

Data Collection & Wrangling

- Data collected from SpaceX official records.

- Cleaned missing values, standardized formats, handled categorical data.

- Features prepared for predictive modeling: Payload Mass, Orbit, Launch Site, Landing Pad, Serial.

EDA & Interactive Visual Analytics Methodology

- Visualized relationships between:
- Flight Number vs Launch Site
- Payload Mass vs Launch Site
- Orbit vs Success Rate
- Used seaborn for scatter and catplots
- Generated interactive maps using FoliumBuilt dashboards with Plotly

Flight Number vs Launch Site

- Plotted Flight Number (x-axis) vs Launch Site (y-axis)
- Hue set to Class (success/failure)
- **Observations:**
  - Launches are clustered by site
  - Some sites have more frequent launches
  - Success rates vary slightly by site

Payload Mass vs Launch Site

- Scatter plot of Payload Mass vs Launch Site
- Hue set to Class (success/failure)
- **Observations:**
  - Certain sites carry heavier payloads more often
  - High payloads occasionally correlate with failed launches
  - Distribution of payload mass differs across sites

Payload Mass vs Orbit

•Scatter plot showing relationship between Payload Mass and Orbit type
•**Observations:**
•Low Earth Orbit (LEO) often carries lighter payloads
•Geostationary transfer orbit (GTO) usually has heavier payloads
•Success rate slightly decreases for very heavy payloads in certain orbits

Flight Number vs Orbit

- Shows which orbit types are used for each flight number
- **Observations:**
- Early flights mostly targeted LEO
- Higher flight numbers have more diverse orbit types
- Certain orbits like GTO and Polar are less frequent but consistent for specific missions

Yearly Launch Success Trend

- Line chart showing average launch success per year
- **Observations:**
  - Success rate has generally increased over the years
  - Early years (2010–2012) had a few failures, later years mostly successful
  - Trend shows SpaceX improving reliability over time

Success Rate by Orbit Type

•Bar chart showing success rate for each orbit type

•**Observations:**
•Low Earth Orbit (LEO) has highest success rate
•Geostationary Transfer Orbit (GTO) slightly lower success rate
•Polar and Sun-synchronous orbits are less frequent but mostly successful

SQL Analysis – Unique Launch Sites

Observations:

- There are multiple launch sites, e.g., CCAFS, KSC, VAFB
- Some sites are used more frequently than others

SQL Analysis – Launch Sites Starting with 'CCA'

Observations:

•CCAFS SLC-40 is the most frequent site starting with CCA
•These sites are key for early SpaceX launches

SQL Analysis – Payload Mass by NASA Boosters

- Total payload mass carried by NASA boosters
- **Observations:**
  - NASA missions tend to have heavier payloads
  - CRS missions (resupply) dominate in total payload mass

SQL Analysis – Average Payload by Booster Version

- Average payload mass for F9 v1.1 booster
- **Observations:**
  - F9 v1.1 handles medium-range payloads
  - Newer booster versions generally carry heavier payloads with higher success rates

SQL Analysis – First Successful Ground Pad Landing

- SQL query using MIN function to find first success date
- **Observations:**
- Ground pad landing success achieved after initial flight tests
- Marks the beginning of reusable booster operations

SQL Analysis – Drone Ship Success

- **Observations:**
  - Medium payload missions show high drone ship landing success
  - Certain boosters are more reliable in this payload range

SQL Analysis – Success vs Failure Count

- Count of successful and failed missions
- **Observations:**
  - Majority of missions are successful
  - Failures mostly in early years or high-risk payloads

SQL Analysis – Max Payload Boosters

- Boosters carrying maximum payload mass

- **Observations:**
- Newer booster versions (F9 B5) carry the heaviest payloads
- These boosters show consistent success even with high mass

SQL Analysis – Failure Outcomes by Month & Year

•Extracted month/year from date column

•**Observations:**

•Failures are scattered across months, mostly in early years
•Some months have multiple failures, indicating testing phases

Folium Map – Launch Sites

Observations:

•Key launch sites clustered in Florida and California
•Map helps visualize geographic spread

Folium Map – Success/Failure

- **Observations:**
- Majority of sites have more successful launches than failures
- Drone ship landings highlighted separately

Folium Map – Distances to Landmarks

Observations:

- Distance analysis useful for logistical planning
- Ground pad sites generally closer to infrastructure

Plotly Dash Dashboard

- Visualizes:
  - Launch success vs Flight Number
  - Payload Mass vs Orbit
  - Launch site statistics
- Interactive filters for dynamic exploration
- **Observations:**
  - Users can explore trends dynamically
  - Dashboard allows for comparison across launch sites and orbits

Predictive Analysis Methodology

- Target: Class (0 = failure, 1 = success)
- Features: Payload Mass, Orbit, Launch Site, Landing Pad, Serial (one-hot encoded)
- Preprocessing: StandardScaler, train/test split (80/20)
- Models: Logistic Regression, SVM, Decision Tree, KNN
- Hyperparameter tuning: GridSearchCV, 10-fold CV

Test Accuracy of Models

Logistic Regression: **0.85**

- SVM: **0.92**
- Decision Tree: **0.88**
- KNN: **0.86**
- Best performing model: **SVM**
- **Observations:**
- SVM often performs best for complex relationships
- Decision Tree is interpretable but may overfit on small data
- KNN is sensitive to feature scaling
- Logistic Regression provides a reliable baseline accuracy

Conclusion

- Successfully explored SpaceX launch data with EDA, SQL, and visual analytics
- Built predictive models with good accuracy
- Developed interactive maps and dashboards for insights
- **Recommendations:**
  - Include more features (e.g., weather, booster reuse)
  - Explore advanced models (Random Forest, XGBoost)
- Future work: Enhance dashboards and predictions for operational insights