Exploratory Data Analysis Report on Data Science Posts

Introduction

This report presents the results of an exploratory data analysis (EDA) conducted on a dataset scraped from an Open Data Science website, which features news and insights related to data science, machine learning, and artificial intelligence. The data was collected using **Selenium**, which enabled automated scraping of relevant posts from the site's **Modeling** category. After collecting the data, several cleaning and preprocessing steps were performed to ensure its quality and structure. This included handling missing values, processing dates, and augmenting features for better analysis.

The cleaned dataset, now free from missing values and inconsistencies, consists of **1,210 posts** with essential attributes such as **headline**, **publisher**, **posting date**, and **category**. This dataset has been delivered, ensuring its quality for further analysis. The data is now ready for additional exploration and analysis, offering valuable insights into trends and content within the data science field.

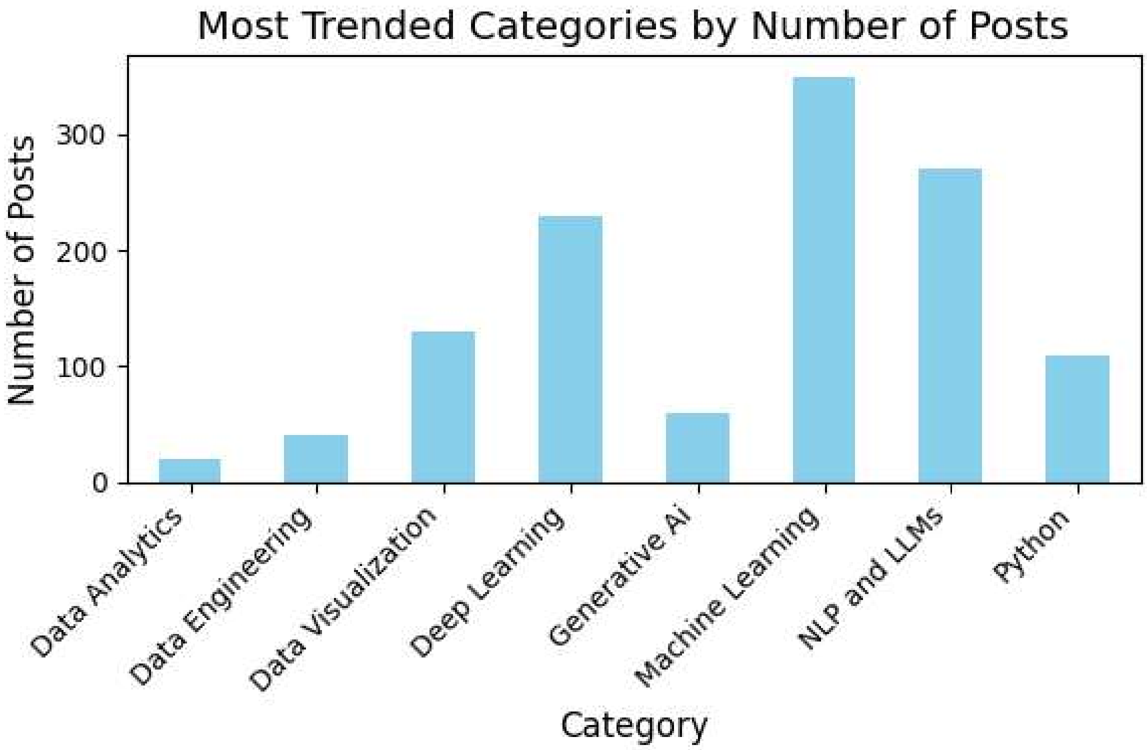
Exploratory Data Analysis (EDA)

This section explores trends and patterns within the dataset by answering key questions through data visualization and analysis.

Q1: What Is the Most Popular and Trending Category in Data Science News?

Answer:

The most trended category on the website is **Machine Learning**, which consistently receives the most posts and remains a dominant topic within the data science community.



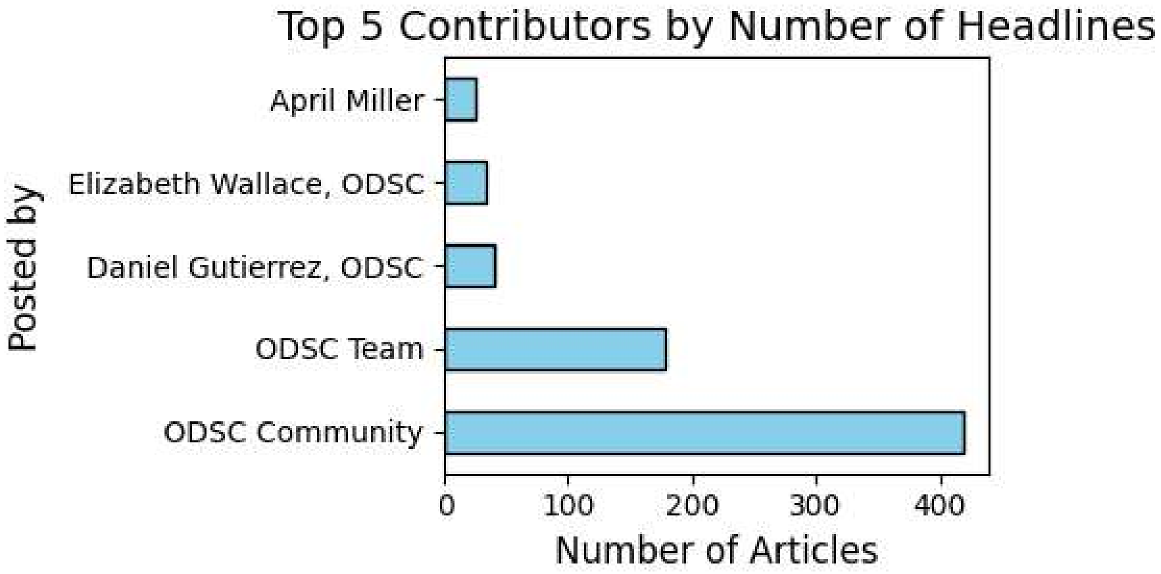
Q2: Who Are the Leading Contributors to Data Science News?

Answer:

The top publishers in terms of post frequency on the site are:

- **ODSC Community**: 418 posts
- **ODSC Team**: 178 posts
- **Daniel Gutierrez, ODSC**: 41 posts
- **Elizabeth Wallace, ODSC**: 33 posts
- **April Miller**: 26 posts

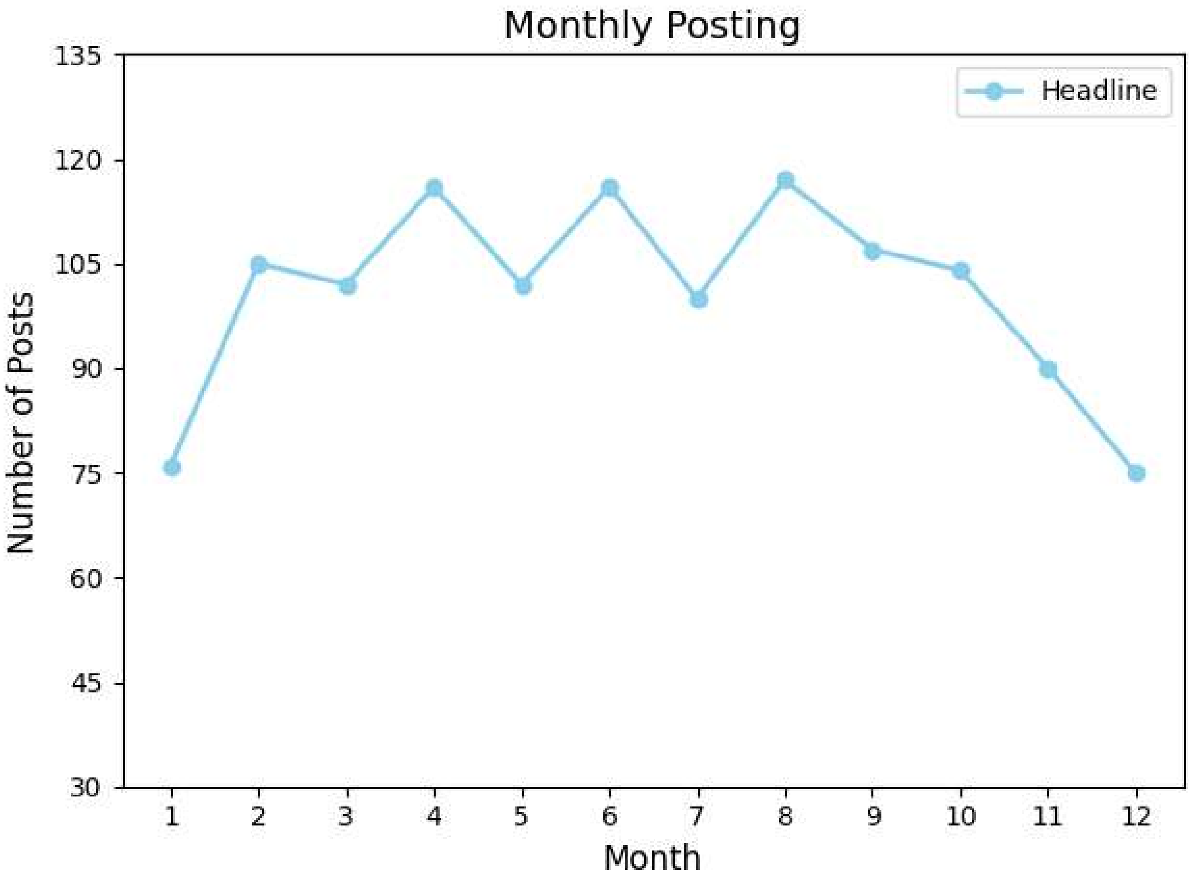
These publishers contribute the most content to the site, with **ODSC Community** as the standout contributor.



Q3: Which Months Have the Highest Post Activity?

Answer:

While there is generally no dramatic fluctuation in posting activity across the months, **April**, **June**, and **August** stand out with noticeably higher posting rates compared to other months. These months appear to see a surge in content creation, which could reflect seasonal trends or key events in the data science community during these periods.



Q4: How Have Data Science Categories Evolved Over the Years?

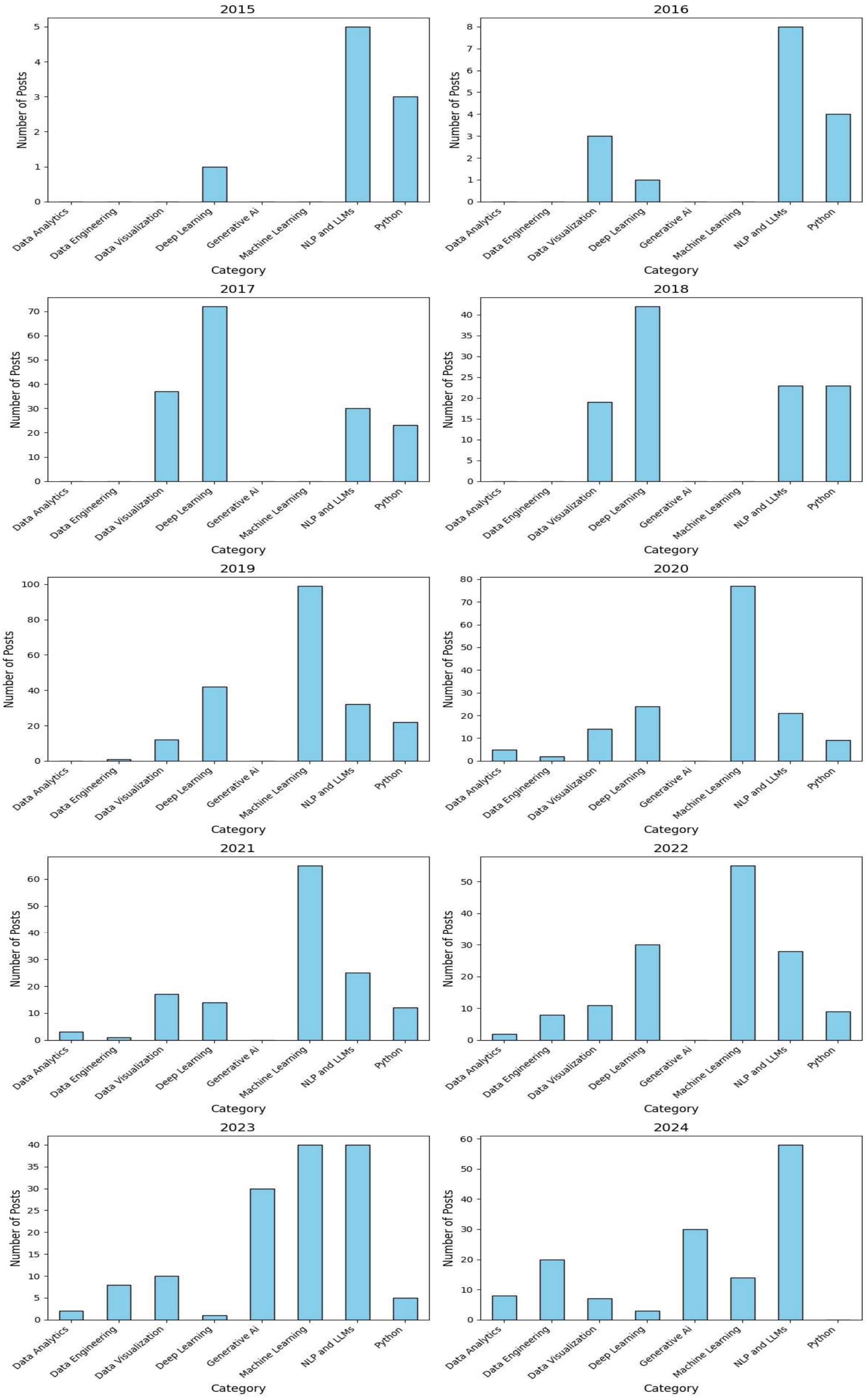
Answer:

The evolution of data science categories over the years reveals a clear upward trend for **Machine Learning**, **NLP and LLMs**, and **Deep Learning**. These fields have gained substantial traction, with **Machine Learning** and **Deep Learning** emerging as the most frequently covered topics on the site.

Category Distribution by Year:

- **Data Analytics**: 20 posts
- **Data Engineering**: 40 posts
- **Data Visualization**: 130 posts
- **Deep Learning**: 230 posts
- **Generative AI**: 60 posts
- **Machine Learning**: 350 posts
- **NLP and LLMs**: 270 posts
- **Python**: 110 posts

Key trends show the growth of **Machine Learning** and **Deep Learning** over time, with these categories receiving a higher number of posts each year.


2015
Number of Posts
0
1
2
3
4
5
Data Analytics
Data Engineering
Data Visualization
Deep Learning
Generative Ai
Machine Learning
NLP and LLMs
Python
Category
2016
Number of Posts
0
1
2
3
4
5
6
7
8
Category
2017
Number of Posts
0
10
20
30
40
50
60
70
Category
2018
Number of Posts
0
5
10
15
20
25
30
35
40
Category
2019
Number of Posts
0
20
40
60
80
100
Category
2020
Number of Posts
0
10
20
30
40
50
60
70
80
Category
2021
Number of Posts
0
10
20
30
40
50
60
Category
2022
Number of Posts
0
10
20
30
40
50
Category
2023
Number of Posts
0
5
10
15
20
25
30
35
40
Category
2024
Number of Posts
0
10
20
30
40
50
60
Category

Data Summary:

- The dataset consists of **1,210 entries** spanning from 2015 to 2024, covering data science news posts.
- **Key columns** in the dataset include:
  - **Headline**: Title of the post
  - **Posted by**: Publisher of the post
  - **Date**: Date when the post was published
  - **Category**: Category the post falls under
- The **cleaned dataset** is complete, with **no missing values** and has undergone rigorous data cleaning to ensure its quality for analysis.

Conclusion

This exploratory data analysis highlights several significant trends in the data science news domain. **Machine Learning** emerges as the most popular category, while **ODSC Community** leads as the most prolific publisher. The analysis also identifies **April**, **June**, and **August** as the months with the highest posting activity and showcases the increasing focus on **Machine Learning** and **Deep Learning** over the years.

The insights from this analysis, based on the cleaned and processed dataset, provide a comprehensive overview of content trends and help identify the most prominent topics and contributors in the data science community.