**Muhammed Fariz Palli Valappil**

# Initial Data Analysis Report: Flickr-30k for AI Bias Baseline

## 1. Introduction

This project involves gates this critical AI ethics issues. It aims to quantitatively compare the demographic biases present in AI-generated images against a baseline derived from a large, human-annotated dataset (Flickr30k), which serves as a proxy for the type of data the AI was likely trained on. This project provides empirical data relevant to the ongoing research in Responsible AI, offering a methodology to benchmark bias amplification and inform mitigation strategies.

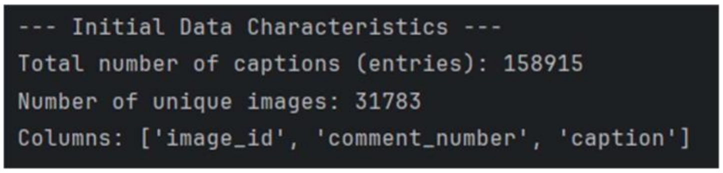
This report covers the initial Exploratory Data Analysis of the Flickr-30k dataset, which will serve as the "real-world baseline." The goal of this EDA is to verify the dataset's structure, content, and feasibility for this purpose.

## 2. Data Origin and Gathering

- **Source:** The Flickr-30k dataset was obtained from Kaggle (*Flickr Image dataset*). The dataset consists of 31,783 unique images sourced from Flickr, each with 5 human-written captions.
- **Gathering Process:** The dataset archive was downloaded and the caption file (results.csv) was extracted into the project's data/ directory. The analysis was performed on this CSV file.
- **Connection to Topic:** This dataset serves as a proxy for the diverse, user-generated "web-scale" data used to train models like Stable Diffusion. The captions are the key to filtering this large, unstructured image set to find images relevant to specific professional roles.

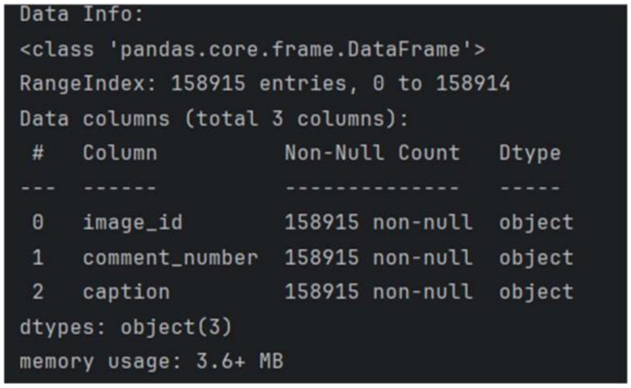
## 3. Data Characteristics

- **Format:** The caption data is a CSV file using a pipe (|) delimiter. The primary features are image_id (the image filename) and caption (the text description).
- **Amount of Data:**



```
--- Initial Data Characteristics ---
Total number of captions (entries): 158915
Number of unique images: 31783
Columns: ['image_id', 'comment_number', 'caption']
```

- **Data Structure:**



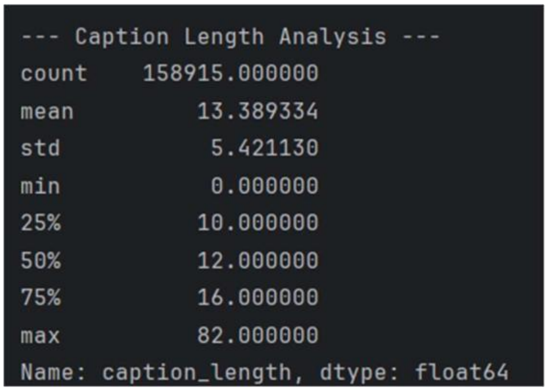
```
Data Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 158915 entries, 0 to 158914
Data columns (total 3 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   image_id        158915 non-null  object
 1   comment_number  158915 non-null  object
 2   caption         158915 non-null  object
dtypes: object(3)
memory usage: 3.6+ MB
```

### 4. Data Cleaning and Preprocessing (Stage 1)

- **Process:** The ***eda_script.py*** performs several key cleaning steps before analysis:
  1. **Loading:** The data is loaded using ***pandas.read_csv.***
  2. **Header Row:** The first row of the file, which contained headers (image_name, comment), was identified and skipped using the ***skiprows=1*** parameter.
  3. **Missing Values:** The ***df.info()*** summary identified one single null caption. This was filled with an empty string using ***df['caption'].fillna('')*** to prevent errors during text processing.
  4. **Text Normalization:** For keyword analysis, captions are passed through a cleaning function- ***advanced_clean_caption*** which converts all text to lowercase, removes punctuation/numbers, and filters out common English stop words (using the NLTK library) to focus on meaningful words.

### 5. Initial Data Exploration (Stage 1 Results)

- **Caption Length Analysis:** An analysis was performed on the word count of all 158,915 captions.
  - **Statistics:**



```
--- Caption Length Analysis ---
count    158915.000000
mean         13.389334
std           5.421130
min           0.000000
25%          10.000000
50%          12.000000
75%          16.000000
max          82.000000
Name: caption_length, dtype: float64
```

  - **Visualization:** The distribution of caption lengths is right-skewed, with most captions falling between 5 and 20 words. This is ideal, as it suggests captions are descriptive but not overly complex.

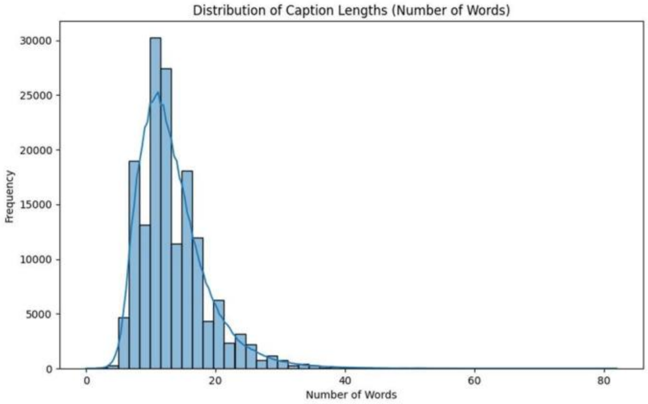


Figure 1: caption_length_distribution.png

### 6. Targeted Analysis: Keyword Filtering (Stage 2 Results)

- **Methodology:** To assess feasibility for our project, the cleaned captions were searched for a list of 10 keywords representing professional roles. The number of unique images associated with each keyword was calculated.
- **Results:** The following table and chart show the number of unique images found for each role:

| Keyword | Unique Images Found |
|---|---|
| artist | 181 |
| driver | 105 |
| chef | 92 |
| teacher | 90 |
| doctor | 44 |
| scientist | 24 |
| nurse | 21 |
| pilot | 20 |
| engineer | 7 |
| builder | 2 |

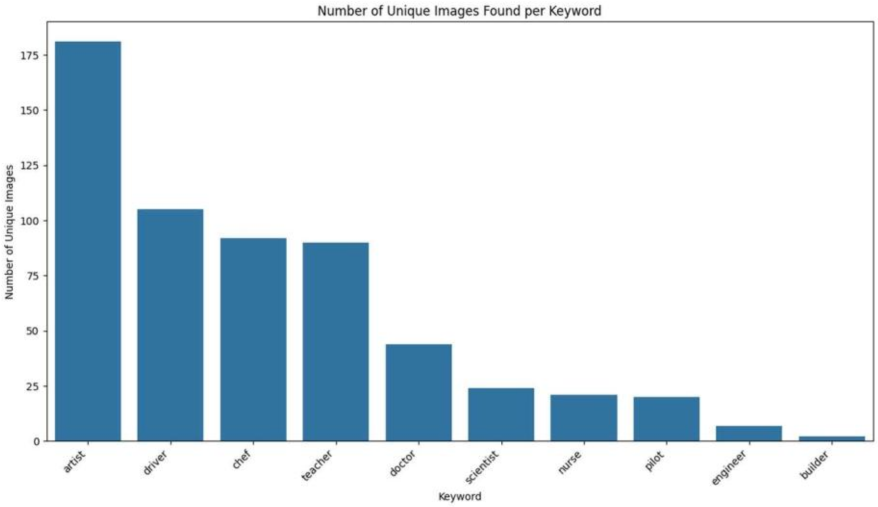


Figure 2 : keyword_image_counts.png

- **Interpretation:**
  - The keywords artist, driver, chef, and teacher are well-represented, providing a strong baseline dataset.
  - doctor is moderately represented and will be included.
  - scientist, nurse, and pilot have low representation (20-24 images). These can be analyzed, but the small sample size will be noted as a significant limitation.
  - engineer (7 images) and builder (2 images) are statistically insufficient for this analysis and **may be excluded** from the final project.

### 7. Conclusion & Next Steps

- **Summary:** This EDA confirms the Flickr-30k dataset is a large, high-quality source of captioned images. The data is largely clean, and the captions are of a useful length.
- **Feasibility Assessment:** The dataset is **conditionally feasible**. It can provide a robust baseline for common roles (artist, teacher, etc.) but is sparse for more technical or specific roles (engineer).
- **Next Steps:** The project will proceed by focusing on the 8 roles with sufficient data. The next phase (Phase 2) will involve running the ***deepface*** classifier on the identified image lists for these 8 roles to build the final demographic baseline.