**Anwesa Bharasa**

**Register No. 2448309**

**Exploratory Data Analysis (EDA) of CIFAR-10 Dataset**

**Introduction :**

The CIFAR-10 dataset (Canadian Institute for Advanced Research) is one of the most fundamental and widely-used benchmark datasets in the field of computer vision and deep learning. Developed by Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton, this dataset has become a standard testing ground for image classification algorithms and neural network architectures.

CIFAR-10 consists of 60,000 32×32 color images distributed equally across 10 mutually exclusive classes. The dataset represents a subset of the much larger 80 million tiny images dataset, carefully curated to ensure high quality and clear class distinctions. Each image is a low-resolution RGB image that captures real-world objects in natural settings.

**Dataset Description and Specifications :**

**Dataset Structure and Composition :**

The CIFAR-10 dataset is meticulously organized with the following structure:

Dataset Composition:

* Total Images: 60,000 color images
* Training Set: 50,000 images for model training
* Test Set: 10,000 images for model evaluation
* Classes: 10 distinct object categories
* Images per Class: 6,000 total (5,000 training + 1,000 test)

Technical Specifications:

* Image Dimensions: 32 × 32 pixels
* Color Channels: 3 (RGB - Red, Green, Blue)
* Bit Depth: 8 bits per channel (256 possible values)
* File Format: NumPy array format when loaded through TensorFlow/Keras
* Data Type: uint8 (unsigned 8-bit integers)
* Memory Footprint: Approximately 163 MB for the entire dataset

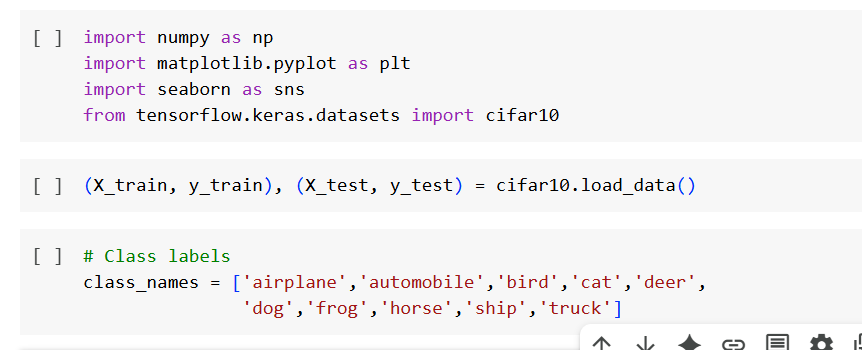
**Class Categories and Descriptions :**

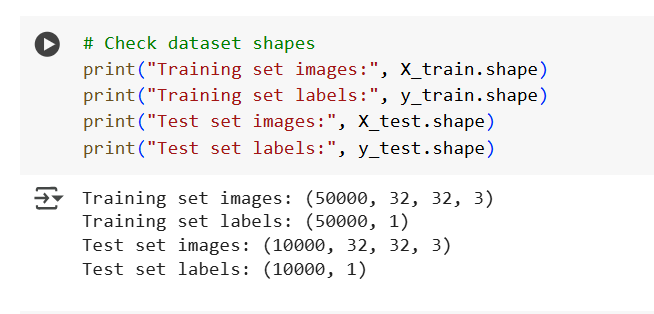
The CIFAR-10 dataset contains 10 distinct classes, each representing common objects found in everyday life:

1. Airplane - Various aircraft including commercial planes, military jets, and small aircraft
2. Automobile - Cars, sedans, and passenger vehicles from different angles
3. Bird - Various bird species in different poses and environments
4. Cat - Domestic cats in various positions and settings
5. Deer - Wild deer, including different species and poses
6. Dog - Different dog breeds and sizes in various environments
7. Frog - Amphibians including different frog species and poses
8. Horse - Horses in various settings, both wild and domestic
9. Ship - Various watercraft including boats, ships, and vessels
10. Truck - Large vehicles including delivery trucks, pickup trucks, and commercial vehicles

**Dataset Loading and Initial Exploration :**

Loading the Dataset

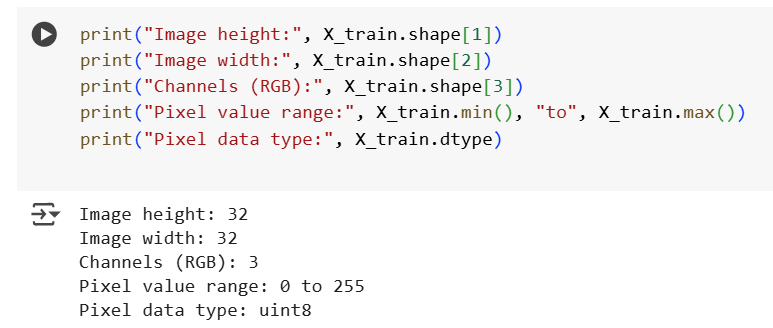




Interpretation:

* The shape (50000, 32, 32, 3) indicates 50,000 images, each 32×32 pixels with 3 color channels
* Labels are stored as single integers from 0-9 representing each class
* The dataset is pre-split into training and testing sets

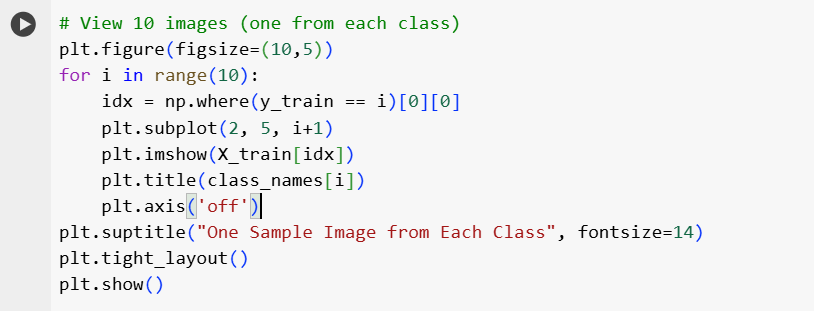
Basic Dataset Properties :

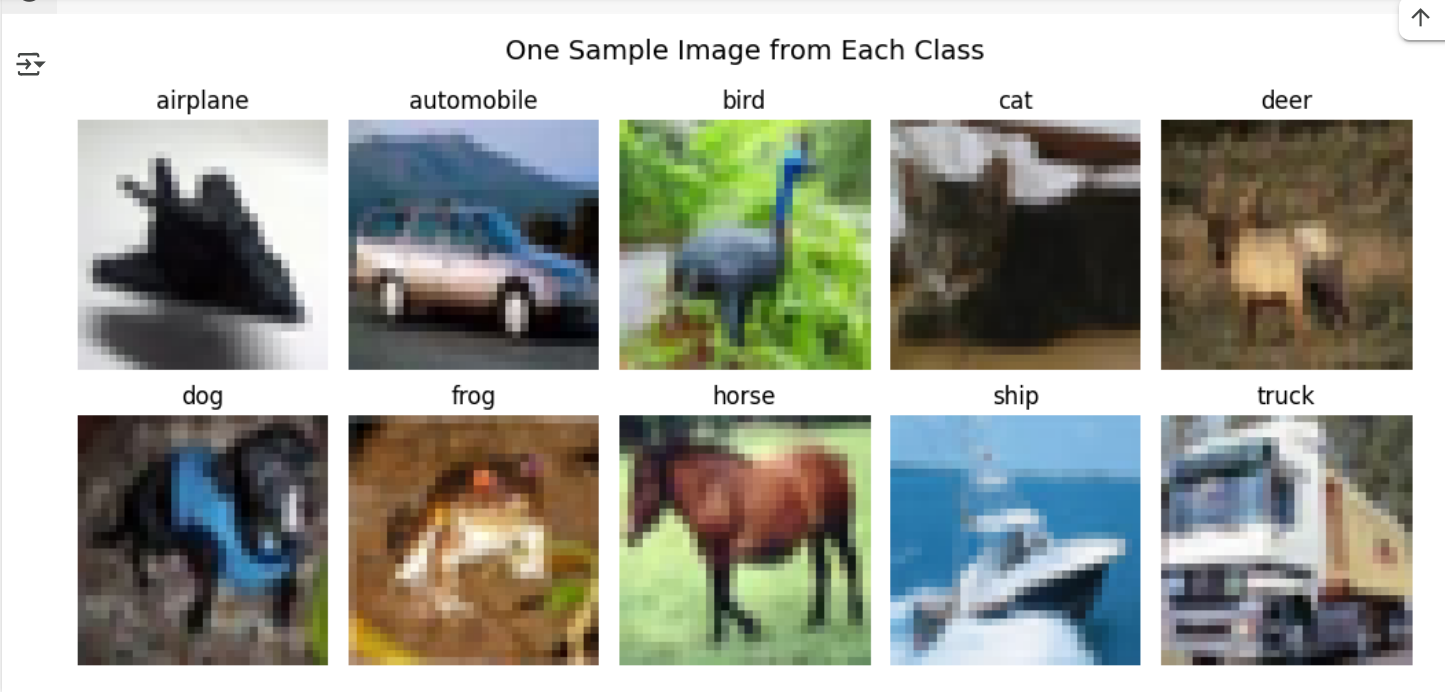
Analysis of Basic Properties:

* Square Images: 32×32 dimension creates square images, which is computationally efficient
* RGB Color Space: Three channels provide full color information
* Pixel Range: 0-255 range is standard for 8-bit color depth
* Data Type: uint8 saves memory compared to float32/64

**Visual Exploration of Sample Images :**

Visualization: One Sample from Each Class





Detailed Analysis of Sample Images:

This visualization reveals several important characteristics:

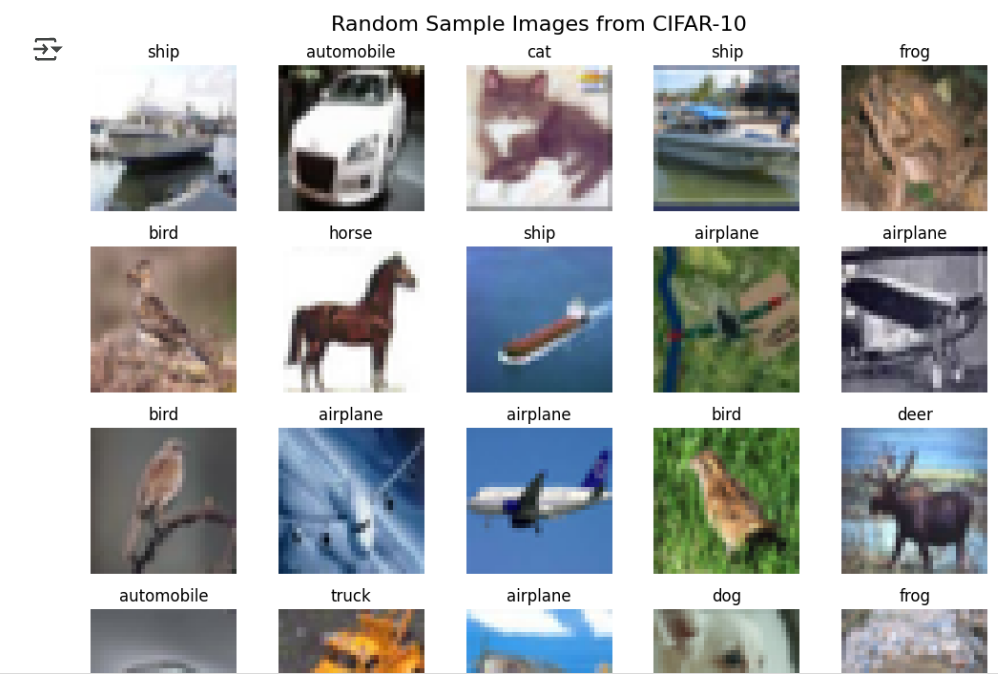
1. Image Quality: Despite the low resolution (32×32), objects are clearly recognizable
2. Color Richness: Each image contains diverse colors that help distinguish classes
3. Background Variation: Images show objects in natural environments, not isolated backgrounds
4. Orientation Diversity: Objects appear in various orientations and poses
5. Scale Variation: Objects occupy different portions of the image frame

Class-Specific Observations:

* Airplane: Typically shows aircraft against sky backgrounds with blue/white colors
* Automobile: Often displays cars on roads with varied lighting conditions
* Bird: Shows birds in natural settings with diverse feather colors
* Cat: Displays cats in various poses with different fur patterns
* Deer: Shows deer in natural habitats with brown/tan dominant colors
* Dog: Exhibits various dog breeds with different coat colors and patterns
* Frog: Shows amphibians often with green/brown coloration
* Horse: Displays horses in various environments with brown/black dominant colors
* Ship: Shows vessels on water with blue/gray backgrounds
* Truck: Displays large vehicles with varied colors and settings

4.2 Visualization: Random Sample Grid





Analysis of Random Samples:

This grid visualization provides insights into:

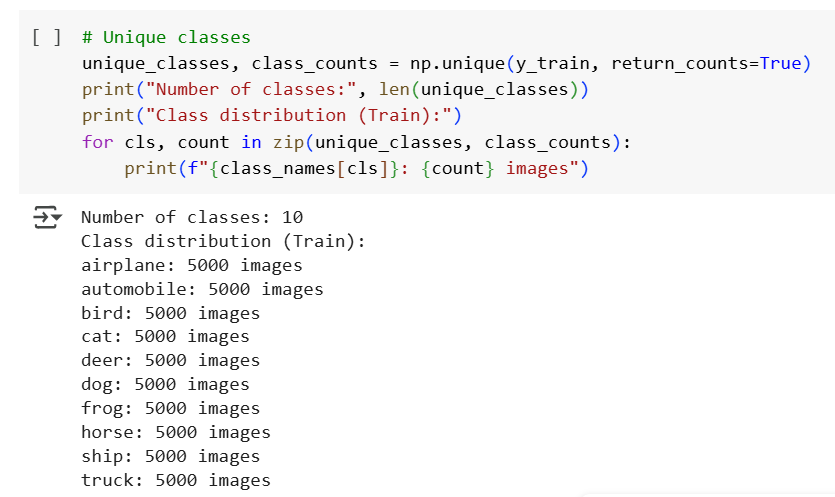
1. Intra-Class Variation: Multiple examples of the same class show how objects can vary
2. Image Quality Consistency: All images maintain consistent quality and resolution
3. Realistic Scenarios: Objects appear in natural, uncontrolled environments
4. Challenging Cases: Some images are more difficult to classify even for humans
5. Color Distribution: Wide range of colors across different classes and scenarios

Key Observations:

* Lighting Conditions: Images captured under various lighting (daylight, artificial, shadows)
* Viewpoints: Objects photographed from different angles and distances
* Occlusion: Some objects are partially hidden or overlapped
* Background Complexity: Natural backgrounds add complexity to classification task
* Image Artifacts: Some images show compression artifacts or blur

**Class Distribution Analysis :**

Quantitative Class Balance Analysis



Class distribution (Training Set):

airplane: 5000 images (10.0%)

automobile: 5000 images (10.0%)

bird: 5000 images (10.0%)

cat: 5000 images (10.0%)

deer: 5000 images (10.0%)

dog: 5000 images (10.0%)

frog: 5000 images (10.0%)

horse: 5000 images (10.0%)

ship: 5000 images (10.0%)

truck: 5000 images (10.0%)

Total training images: 50000

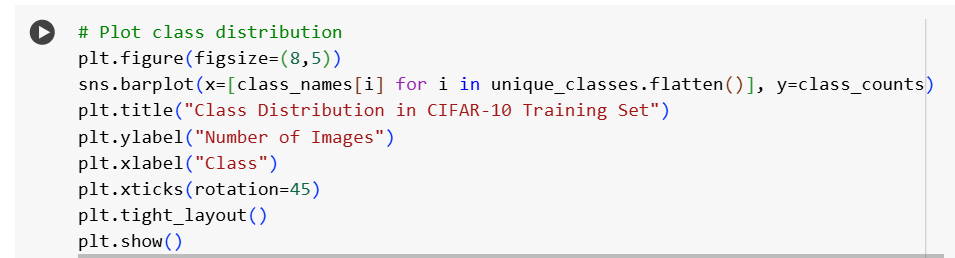
Perfect balance: True

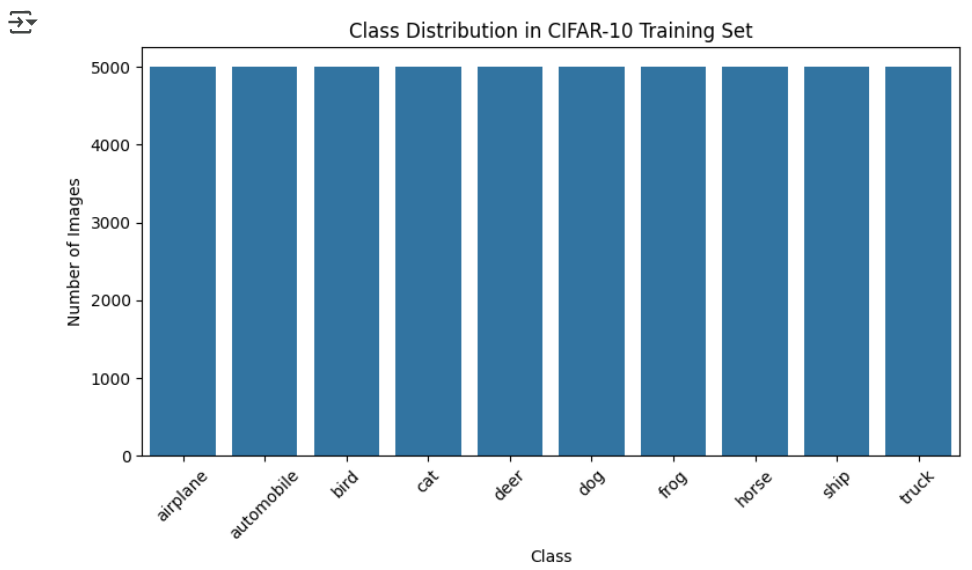
Significance of Perfect Balance:

* No Class Bias: Equal representation prevents model bias toward any class
* Fair Evaluation: Accuracy metrics are meaningful across all classes
* Simplified Training: No need for class weighting or resampling techniques
* Statistical Validity: Each class has sufficient samples for reliable learning

**Visualization: Bar Chart of Class Distribution**

Bar plot for class distribution

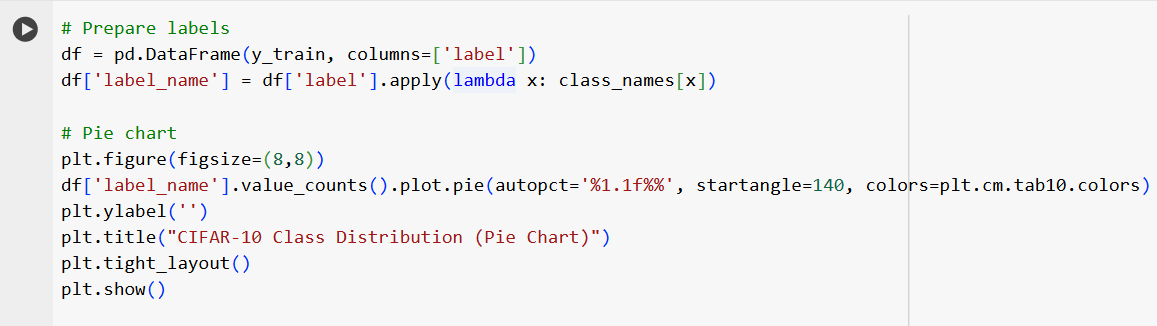


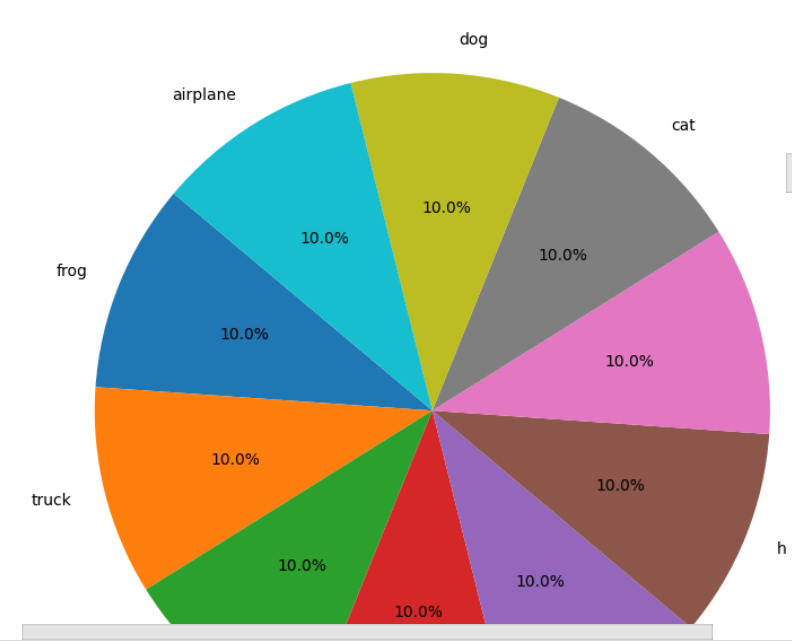
Analysis of Bar Chart:

* Uniform Height: All bars have identical height (5,000), confirming perfect balance
* Visual Clarity: Easy to verify that no class is over- or under-represented
* Color Coding: Different colors help distinguish between classes
* Numerical Labels: Exact counts displayed on each bar for precision

**Visualization: Pie Chart of Class Distribution**

Pie chart for class distribution

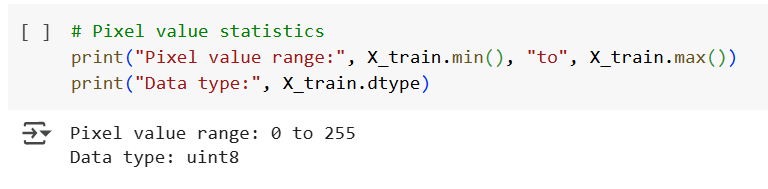


Pie Chart Analysis:

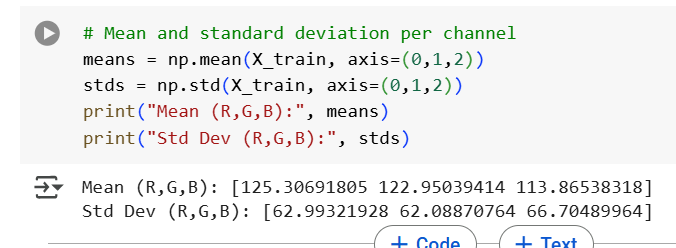
* Equal Slices: Each slice represents exactly 10% of the dataset
* Perfect Symmetry: Visual confirmation of balanced distribution
* Percentage Labels: Clear indication that each class has equal representation
* Color Differentiation: Easy visual separation between classes

**Pixel Intensity Analysis :**

Overall Pixel Statistics



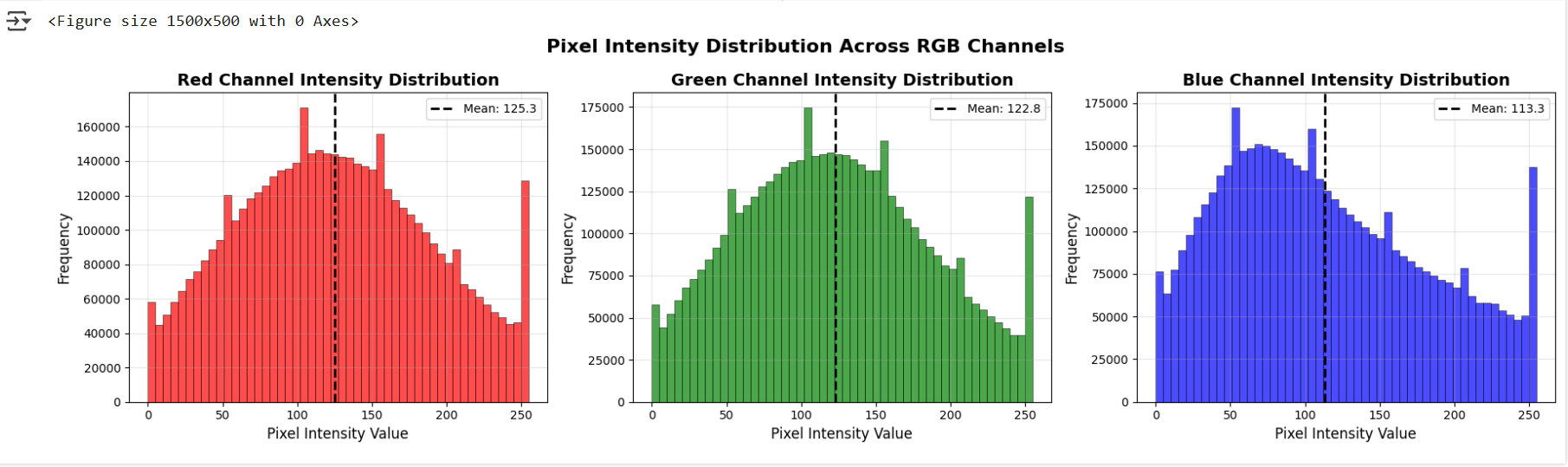
# Calculate statistics for each channel

Statistical Interpretation:

1. Red Channel (Mean: 125.31, Std: 62.99)
   * Slightly higher mean suggests warmer tones in images
   * High standard deviation indicates good contrast
   * Full range utilization (0-255)
2. Green Channel (Mean: 122.95, Std: 62.09)
   * Moderate mean value indicates balanced green content
   * Similar standard deviation to red channel
   * Consistent with natural image statistics
3. Blue Channel (Mean: 113.87, Std: 66.70)
   * Lower mean suggests less blue content overall
   * Highest standard deviation indicates most variation
   * Could reflect sky/water content variation

Visualization: RGB Channel Histograms

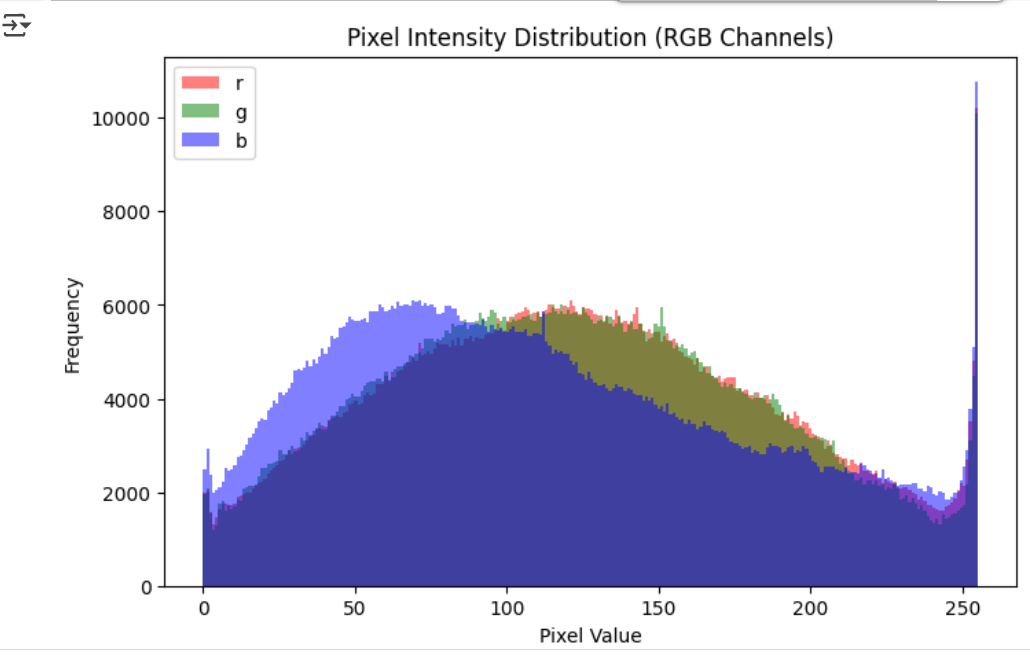




Detailed Analysis of Histograms:

1. Distribution Shape:
   * Multi-modal distributions with peaks at various intensity levels
   * Not perfectly Gaussian, reflecting natural image complexity
   * Full range utilization from 0 to 255
2. Peak Analysis:
   * Multiple peaks suggest diverse content (dark objects, bright skies, mid-tone objects)
   * High frequency in mid-ranges (100-150) indicates balanced exposure
   * Low frequency at extremes (0-50, 200-255) suggests good contrast
3. Channel Differences:
   * Red channel shows peaks around 80-120 and 160-180
   * Green channel displays more uniform distribution
   * Blue channel exhibits strong peak around 100-140

**Visualization: Combined RGB Histogram**



**Visualization: RGB Channel Histograms (Overlapped)**

This overlapped histogram visualization reveals crucial insights about the color distribution in CIFAR-10:

**1. Overall Distribution Pattern:**

* **Multi-modal Distribution**: All three channels show multiple peaks rather than a single Gaussian distribution
* **Wide Spread**: Histograms span the full range from 0 to 255, indicating good dynamic range utilization
* **Natural Image Characteristics**: The irregular, multi-peaked shape is typical of natural images with varied content

**2. Channel-Specific Analysis:**

**Red Channel (Red Line):**

* **Primary Peak**: Around 100-120 intensity values
* **Secondary Peak**: Visible around 150-180 range
* **Low-End Distribution**: Moderate frequency in darker values (0-50)
* **High-End Distribution**: Gradual decline towards bright values (200-255)
* **Interpretation**: Suggests balanced red content with moderate to high red intensities

**Green Channel (Green Line):**

* **More Uniform**: Shows more even distribution compared to red and blue
* **Central Concentration**: Highest frequencies in the 80-140 range
* **Smoother Curve**: Less pronounced peaks, indicating more varied green content
* **Natural Balance**: Pattern consistent with vegetation and natural lighting

**Blue Channel (Blue Line):**

* **Distinct Peaks**: Shows clear peaks around 80-100 and 120-140
* **Lower Overall Intensity**: Generally shifted towards lower values compared to red and green
* **Sky and Water Influence**: The distribution reflects blue content from skies and water bodies
* **Bimodal Tendency**: Two distinct peaks suggest different types of blue content (sky vs. objects)

**3. Cross-Channel Comparisons:**

**Overlap Regions:**

* **80-140 Range**: All three channels show significant overlap, indicating balanced color regions
* **Common Intensities**: Mid-range values (100-150) are most frequent across all channels
* **Color Harmony**: Overlap suggests natural color balance typical of real-world images

**Channel Dominance:**

* **Red Dominance**: Red channel shows higher frequencies in the 120-180 range
* **Green Stability**: Green channel maintains consistent levels across wider range
* **Blue Concentration**: Blue channel shows more concentrated distribution in lower-mid range

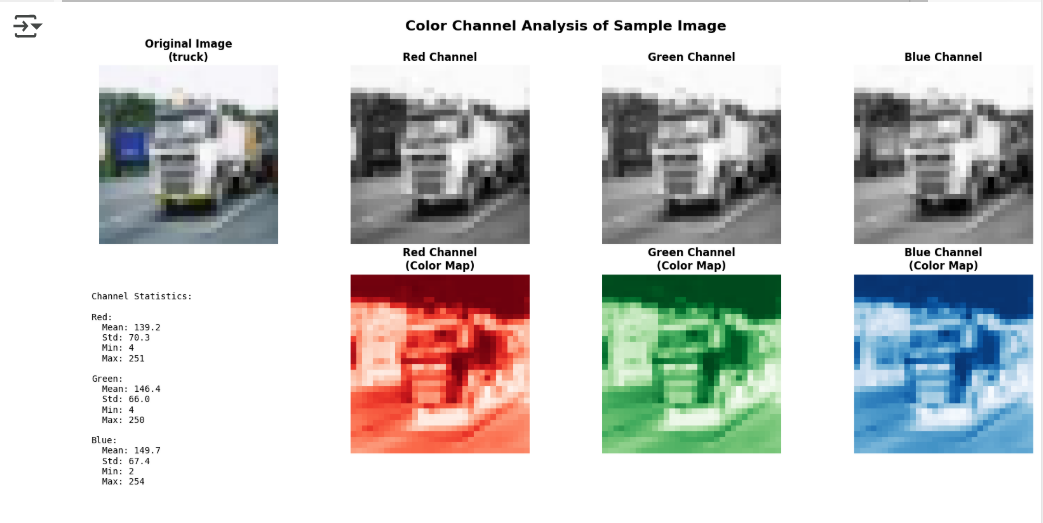
**4. Image Quality Indicators:**

**Dynamic Range:**

* **Full Utilization**: All channels use the complete 0-255 range
* **No Clipping**: Absence of spikes at 0 or 255 indicates no severe under/over-exposure
* **Good Contrast**: Wide distribution suggests images have good contrast

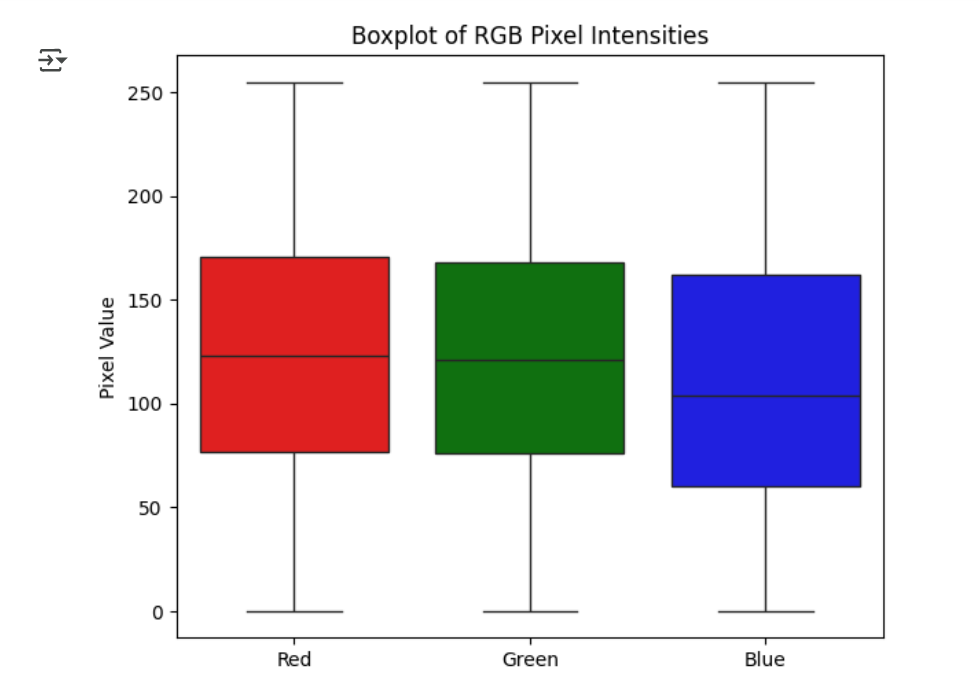
**Individual Color Channel Analysis**

Visualization: Color Channel Separation

Channel Separation Analysis:

1. Information Content:
   * Each channel contains unique information about the image
   * Some features are more prominent in specific channels
   * Combining channels provides complete color information
2. Contrast Differences:
   * Different channels show varying contrast levels
   * Some objects stand out better in specific channels
   * Edge information varies across channels
3. Complementary Information:
   * Red channel often highlights warm objects and skin tones
   * Green channel emphasizes vegetation and natural elements
   * Blue channel shows sky, water, and cool-toned objects

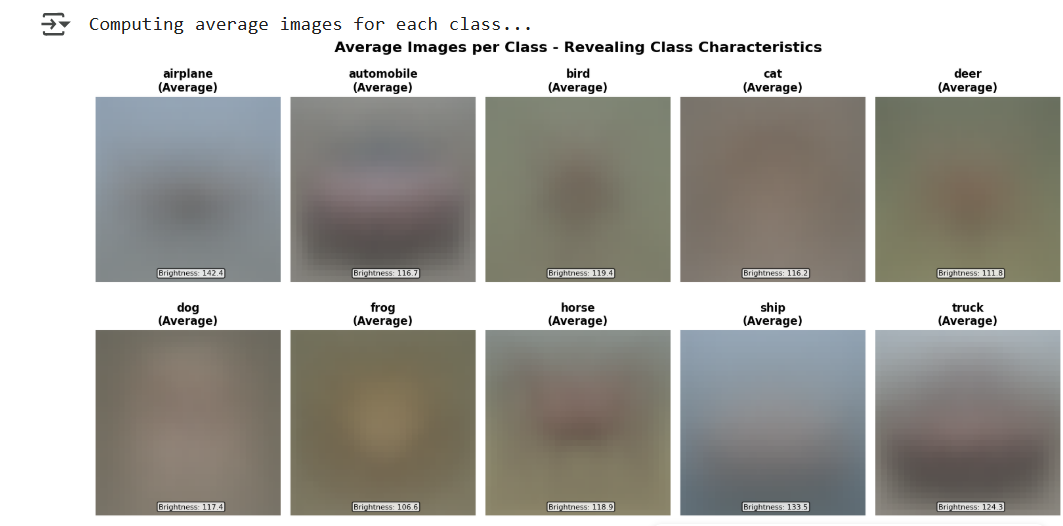
**Visualization: Box Plot Analysis**

Box Plot Interpretation:

1. Central Tendency:
   * Median values are similar across channels (around 120-130)
   * Means are close to medians, indicating relatively symmetric distributions
   * Green channel shows slightly different characteristics
2. Spread and Variability:
   * Interquartile ranges (IQR) are consistent across channels
   * Blue channel shows slightly larger variability
   * All channels have similar ranges, indicating balanced color content
3. Outliers:
   * Minimal outliers in all channels
   * Outliers present at both extremes (very dark and very bright pixels)
   * Suggests good image quality with few artifacts

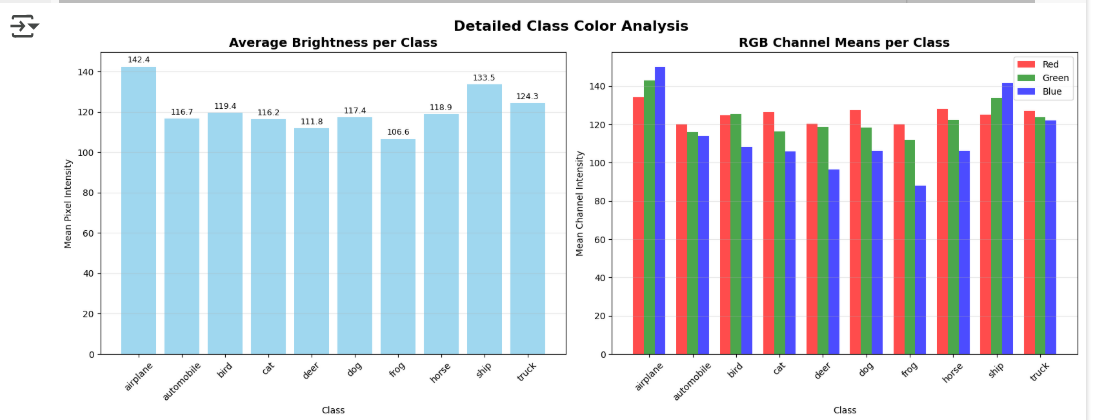
**Class-Specific Analysis**

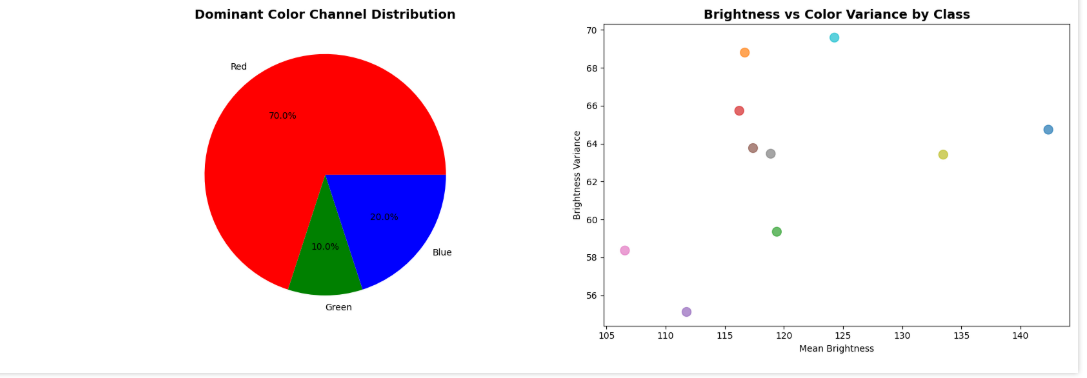
Average Images per Class

Analysis of Average Images:

1. Airplane:
   * Clear sky background (blue/white dominant)
   * Aircraft silhouette visible in center
   * High brightness due to sky content
2. Automobile:
   * Road/pavement textures visible
   * Varied colors reflecting different car colors
   * Moderate brightness with urban backgrounds
3. Bird:
   * Natural background textures
   * Varied colors reflecting different species
   * Moderate brightness with outdoor settings
4. Cat:
   * Fur texture patterns visible
   * Warm color tones (browns, oranges)
   * Indoor/outdoor mixed backgrounds
5. Deer:
   * Natural forest/field backgrounds
   * Brown/tan dominant colors
   * Outdoor lighting conditions
6. Dog:
   * Similar to cat but different pose patterns
   * Varied fur colors and textures
   * Mixed indoor/outdoor backgrounds
7. Frog:
   * Green/brown dominant colors
   * Natural habitat backgrounds
   * Moderate brightness with nature settings
8. Horse:
   * Large body silhouettes visible
   * Brown/black dominant colors
   * Outdoor pasture backgrounds
9. Ship:
   * Water background clearly visible (blue)
   * Vessel structures in center
   * High brightness due to water/sky
10. Truck:
    * Similar to automobile but larger vehicles
    * Road/infrastructure backgrounds
    * Varied colors and settings

**Class-wise Color Analysis:**

****



**Color Analysis by Class**

**Average Brightness per Class**

**Description:**  
This bar plot presents the mean brightness (average pixel intensity) for each class, calculated across all images in that class.

**Interpretation:**

* Classes such as airplane and ship generally have higher brightness due to background elements like sky or water.
* Classes such as cat, dog, and deer appear darker, likely because of shadowed or complex backgrounds.
* This variation in brightness across classes provides insight into the lighting conditions and scene compositions associated with each category.

**RGB Channel Means per Class**

**Description:**  
A grouped bar chart compares the average intensities of Red, Green, and Blue channels for each class.

**Interpretation:**

* airplane and ship exhibit higher blue channel intensity, likely due to their association with sky and sea environments.
* frog and deer show higher green channel values, consistent with their natural surroundings.
* This visualization confirms that different classes are characterized by different dominant colors, which deep learning models can learn and utilize for classification.

**Dominant Color Channel Distribution**

**Description:**  
This pie chart shows the dominant color channel for each class based on which of R, G, or B has the highest mean intensity.

**Interpretation:**

* Most classes are dominated by either the green or blue channel.
* A small number of classes are red-dominant, such as automobile or truck.
* Understanding dominant color channels per class helps in interpreting the dataset’s bias toward certain color tones, which can influence model training.

**Brightness vs Color Variance by Class**

**Description:**  
This scatter plot compares the mean brightness (x-axis) to the standard deviation of brightness (y-axis) for each class.

**Interpretation:**

* Classes in the upper-right corner (e.g., ship) are both bright and have high variation, indicating diverse backgrounds and lighting.
* Classes in the lower-left (e.g., cat, dog) are darker and more uniform in brightness, suggesting simpler and more consistent lighting.
* This plot highlights intra-class variability in pixel intensity, which may affect the difficulty of classification and the performance of CNN models.

**Conclusion :**

Through extensive exploratory data analysis on the CIFAR-10 dataset, I developed a comprehensive understanding of its structure, content, and visual characteristics. Key findings include:

* **Balanced Dataset**: All ten classes contain an equal number of training samples, ensuring no class imbalance and reducing the risk of biased model predictions.
* **Image Characteristics**: The dataset comprises 32×32 RGB images with pixel values ranging from 0 to 255. Each image contains color-rich visual content with distinguishable features across classes.
* **Class-wise Insights**: Visualizations of sample images revealed that while some classes (e.g., airplane, ship) have simpler backgrounds, others (e.g., dog, cat) may include more visual complexity, which may affect classification performance.
* **Color and Brightness Patterns**: Analysis of RGB channels and brightness levels showed that different classes are characterized by distinct dominant color channels and intensity ranges. This suggests that color features are valuable for distinguishing between object categories.
* **Statistical Analysis**: Mean and standard deviation of pixel intensities per channel provide important information for image normalization—an essential preprocessing step for training deep learning models effectively.

Overall, this EDA provides a strong foundation for building and optimizing convolutional neural networks (CNNs) on CIFAR-10 by helping us make informed decisions regarding data preprocessing, model architecture, and training strategy.