Exploratory Data Analysis (EDA) Report

1. **Distribution of Numerical Features (Histograms):**
    - **Stroke (target):** Strong class imbalance with most cases being non-stroke (0), might have biased models favoring the majority class.
    - **Hypertension & Heart Disease (binary):** Highly skewed toward 0, limiting generalization in detecting positive cases.
    - **BMI:** Nearly normal with right-skewed outliers, which could affect models sensitive to variance.
    - **Avg_glucose_level:** Bimodal distribution (~100 and ~150), potentially indicating different health groups. Bimodality could confuse models without normalization.
    - **Age:** Fairly uniform across younger and middle-aged groups but slight underrepresentation of the elderly may affect age-sensitive model performance.
2. **Distribution of Categorical Features (Count Plots):**
    - **Smoking_status:** Most individuals have never smoked, potentially limiting analysis into smoking-related stroke risks.
    - **Work_type:** Dominance of private-sector workers, underrepresentation in "never worked" may distort work-related analyses.
    - **Residence_type:** Balanced urban-rural split, suitable for analysis by residence.
    - **Marital_status:** Skewed towards married individuals, may have a complex relation with factors such as age, and gender, and thus the isolation of its impact can be more difficult.
3. **Correlation Heatmap:**
    - **Age & Hypertension (0.28)** and **Age & Heart Disease (0.26):** Indicate age as a risk factor for these conditions.
    - **Age & BMI (0.33):** Weak positive correlation, suggesting a slight BMI increase with age.
    - **Hypertension & Heart Disease (0.11):** Mild trend where hypertension may co-occur with heart disease.
    - **Stroke Correlations:** Weak correlations with stroke (highest is age at 0.25), indicating limited predictive power of individual features.
4. **Box Plots (Numerical Features vs. Stroke):**

- **Hypertension & Heart Disease (binary):** stroke cases tend to have more instances having hypertension and heart disease. However, these features are not strong differentiators due to predominant 0 values.

- **Gender:** Balanced across stroke groups, unlikely to be a strong predictor in isolation.

- **BMI:** The distribution between stroke and non-stroke groups is quite similar. Outliers could introduce noise, may need handling strategies (scaling/removal).

- **Avg_glucose_level:** Higher glucose levels appear more common in the stroke group, suggesting a potential link but this might be hindered by outliers in non-stroke cases.

- **Age:** There is a strong distinction between stroke and non-stroke cases, with stroke patients tending to be older.

Key Challenges of the Dataset:

- **Class Imbalance**: Stroke cases are rare, leading models to favor the majority non-stroke class, potentially reducing accuracy for stroke predictions.
- **Categorical Feature Imbalance:** Skewed distributions in smoking_status and work_type may limit insights from less common categories.
- **Weak Correlation with Stroke**: Features like age, hypertension, and heart_disease have low correlations with stroke, suggesting limited predictive strength when used individually.
- **Outliers and Variance:** Outliers in BMI and avg_glucose_level could skew model performance, requiring normalization or mitigation.
- **Bimodal Distribution:** avg_glucose_level shows bimodality, complicating analysis without proper handling of distinct health groups.
- **Moderate Multicollinearity:** Moderate correlations between features (e.g., age with hypertension and heart_disease) may cause redundancy, impacting model interpretation.