Assignment 1

Name : Lokesh Mamillapalli

Roll No : 2023101115

Name : Nethavath Praveen

Roll No : 2023102013

Assignment Done on Kaggle NoteBooks

Part 1 link : https://www.kaggle.com/code/nethavathpraveen/data-analytics-assignment1-part1

Part 2 Link : https://www.kaggle.com/code/nethavathpraveen/data-analytics-assignment1-part2

M25 Data Analytics I Assignment 1.pdf

Part - 1 Report

Data Cleaning Methodology

The following steps were performed based on the analysis in the provided notebook:

- **Handling Missing Values**: The dataset contained missing values in the 'Workclass', 'Occupation', and 'Country' columns. These were handled by filling the missing entries with the **mode** (the most frequently occurring value) of their respective columns. This is a standard approach to preserve the dataset's size without significantly skewing the distributions.
- **Feature Engineering**: New columns were created to provide more analytical power:
  - `Net_Capital`: Created by subtracting `Capital_Loss` from `Capital_Gain` to provide a single metric for an individual's overall capital activity.
  - `Any_Capital_Activity`: A binary flag (0 or 1) was created to easily identify individuals with any non-zero capital gains or losses.
- **Data Transformation and Grouping**: To improve interpretability and prepare the data for modeling, several continuous and high-cardinality features were grouped into broader categories:
  - **Education Group**: The 16 unique education levels were consolidated into 5 distinct categories: **Elementary, Secondary, College/Diploma, Graduate, and Postgraduate**.
  - **Age Group**: The continuous 'age' variable was binned into three groups: **Young (17-30), Middle-aged (31-50), and Senior (51-90)**.
  - **Work Intensity**: 'Hours per week' was categorized into **Part-time (<35 hrs), Full-time (35-45 hrs), and Overtime (>45 hrs)**.

Education Distribution & Grouping

The initial analysis of the dataset revealed **16 unique education levels**. The distribution was heavily skewed towards certain categories, with **'HS-grad' (15,784 individuals)**, **'Some-college' (10,878 individuals)**, and **'Bachelors' (8,025 individuals)** being the most

frequent. To simplify this feature and make it more effective for analysis and modeling, these 16 levels were consolidated into five broader, more meaningful categories:

- **Elementary**: Grades 1st-4th, 5th-6th, 7th-8th, 9th, 10th, 11th, 12th
- **Secondary**: HS-grad
- **College/Diploma**: Some-college, Assoc-acdm, Assoc-voc, Prof-school
- **Graduate**: Bachelors
- **Postgraduate**: Masters, Doctorate

The reasoning behind this grouping was to create distinct, ordered categories that represent a clear progression in educational attainment. After grouping, **Secondary** education became the largest category with 32.3% of the population, followed by **College/Diploma** at 27.6%. This categorization makes the educational landscape of the dataset much easier to interpret at a glance.

Age–Work Intensity Relationship and Grouping

The distributions of both 'age' and 'hours-per-week' were examined to understand the demographic and work patterns.

- **Age**: The age distribution was right-skewed, with the majority of individuals falling between the ages of 25 and 45. To facilitate analysis, age was grouped into three categories:
  - **Young**: 17-30 years
  - **Middle-aged**: 31-50 years
  - **Senior**: 51-90 years
  - The **Middle-aged** group was the largest, comprising nearly half of the dataset.
- **Work Intensity**: The 'hours-per-week' distribution showed a significant peak at 40 hours, which is the standard for full-time work. This variable was grouped as follows:
  - **Part-time**: < 35 hours
  - **Full-time**: 35-45 hours
  - **Overtime**: > 45 hours
  - **Full-time** workers were the most common group.

Analysis of the relationship between the grouped 'Age' and 'Work Intensity' variables revealed that the **Middle-aged** group had the highest proportion of individuals working **Overtime**. In contrast, the **Young** and **Senior** groups had higher proportions of **Part-time** workers. Grouping these variables made these trends much clearer and more interpretable than when analyzing the raw, continuous data.

Capital Gains/Losses and Group Performance

The analysis of 'capital-gain' and 'capital-loss' showed that a vast majority of individuals (**over 91%**) had no capital activity (i.e., both gain and loss were zero). For the minority with non-zero values, the distributions were highly skewed.

To analyze performance, a **'Net_Capital'** feature was created (Capital Gain - Capital Loss). When comparing the average 'Net_Capital' across the previously defined groups, it was found that:

- **Age**: The **Senior** group had the highest average net capital, suggesting that capital gains tend to increase with age.
- **Work Intensity**: The **Overtime** group showed a higher average net capital compared to the 'Full-time' and 'Part-time' groups.

The analysis indicated that **age has a stronger and more consistent association with net capital** than work intensity. As age increases, both the likelihood of having any capital activity and the average net capital amount tend to rise.

Final Dataset Refinement and Structure

The dataset was refined to improve its structure and readiness for modeling. The key changes include:

- **New Features Created**:
  - `education_group`: Categorical feature for grouped education levels.
  - `Age_Group`: Categorical feature for grouped age.
  - `Work_Intensity`: Categorical feature for grouped hours per week.
  - `Net_Capital`: Numerical feature representing `Capital_Gain` - `Capital_Loss`.
  - `Any_Capital_Activity`: A binary feature indicating whether an individual had any capital gain or loss.
- **Features Removed**: No features were explicitly removed in the analysis provided, but for modeling purposes, the original 'education', 'age', and 'hours-per-week' columns could be replaced by their grouped counterparts.
- **Category Counts**: Grouping significantly reduced the number of unique categories for 'education' (from 16 to 5), 'age' (from 74 to 3), and 'hours-per-week' (from 96 to 3), which helps in reducing noise and improving model performance.
- **Missing Values**: The data cleaning process handled missing values in the original features. The creation of new grouped features did not introduce any new missing values, ensuring data quality was maintained.

Part - 2 Report

Data Cleaning Methodology

1. Missing Value Treatment

- **High Missing Columns**: Removed columns with >80% missing values to improve data quality
- **Price Cleaning**: Converted price to numeric format and imputed missing values using city-wise averages
- **Carpet Area**: Standardized numeric format with city-based imputation for missing values
- **Categorical Variables**: Filled missing values with appropriate defaults (e.g., 'Unknown' for developers)

2. Data Standardization

- **Column Names**: Stripped whitespace and standardized naming conventions
- **Date Processing**: Converted availability dates to datetime format for temporal analysis
- **Amenity Conversion**: Transformed all amenity columns to binary format (0/1) for consistent analysis

3. Outlier Management

- **IQR Method**: Applied Interquartile Range technique to remove price and area outliers

1.Price Segmentation & Market Overview

- **Market Structure**: Properties segmented into Low Budget (<33rd percentile), Medium Budget (33rd-67th percentile), and High Budget (>67th percentile)
- **Property Type Patterns**: Apartments constitute the majority across all budget categories
- **Amenity Availability**: Higher-budget properties show significantly better amenity packages

2.City-Level Comparative Analysis (Mumbai vs Thane)

- **Price Premium**: Mumbai commands 2-3x higher prices than Thane across comparable property types
- **Space Efficiency**: Thane offers larger carpet areas at lower price points
- **Market Composition**: Residential properties dominate both cities (>85%)
- **Commercial Opportunities**: Limited but concentrated in Mumbai's premium segments

Investment Insight: Mumbai offers stability and premium positioning; Thane provides growth potential and value for money.

3.Location-Based Premium Analysis

- **Prime Location Premium**: Properties located in **prime areas** are priced about **15–25% higher** compared to similar properties in non-prime areas
- **Amenity Justification**: Prime locations offer 30-40% better amenity packages
- **Value Proposition**: Location premium justified by brand value and facilities rather than space

Investment Insight: Prime location investments command premiums but also offers better amenities and brand value.

4.Value-for-Money Opportunities

- **Best Value Cities**: Thane leads in carpet area per rupee invested
- **Optimal Segments**: Budget and mid-range properties offer best value scores
- **Configuration Efficiency**: 2-3 BHK apartments provide optimal space-price balance
- **Investment Strategy**: Focus on non-prime Thane properties for maximum value

5.Feature & Amenity Impact on Price

- **High-Impact Amenities**: Skydeck, Sea Facing, Sky Villa show highest price correlation
- **Essential Features**: Swimming Pool, Gymnasium, Club House are market standards
- **City Preferences**: Mumbai values luxury amenities more than Thane
- **ROI Factors**: Premium amenities lead to a 20–35% increase in prices

6.Timeline & Readiness Effect on Pricing

- **Ready-to-Move Premium**: 10-15% more costly over under-construction properties
- **Risk-Return**: Under-construction offers higher appreciation potential but with delivery risks
- **Construction Timeline**: 2024-2026 shows peak supply across both cities
- **Market Timing**: Future projects (2027+) show price stability

Investment Insight: Short-term investors should prefer ready-to-move; long-term investors can benefit from under-construction projects.

7.Developer Impact on Properties

- **Brand Premium**: Top-tier developers command 2-3x higher median prices

- **Quality Consistency**: Premium developers show less price variability
- **Amenity Correlation**: Established developers offer significantly better facility packages
- **Market Positioning**: There is a clear distinction between budget, mid-tier, and luxury developers

Investment Insight: Developer selection significantly impacts property value, amenity quality, and future appreciation potential.

Conclusion

The real estate market has different opportunities across cities. Mumbai is stable and better for premium investors, while Thane offers more growth and is good for budget-friendly buyers. To succeed, investors should choose the right developer, invest at the right stage of construction, and match their strategy with their risk and return goals.

Contributions :
Part I: 1,2 and Part II -1,2,3,4 Praveen
Part I: 3,4 and Part II -5,6,7 Lokesh