Phase 1 Documentation – Group 20

Project Title: Remote Work and Urban Traffic Reduction

Course: Introduction of Data Science

Group: 20

Members:

Amina Kainat (SP23-BCS-018)

Fizza Ali (SP23-BCS-043)

Laiba Ajmal (SP23-BCS-060)

Zainab Naeem (SP22-BCS-179)

1. Dataset Description:

This project explores the relationship between remote work adoption and urban traffic congestion reduction. Two datasets were used:

1. Traffic Dataset (TomTom Traffic Index): Contains city-level traffic congestion metrics before and after remote work adoption.
- city: City name
- year: Year of observation
- pre_remote_congestion: Congestion before remote work
- post_remote_congestion: Congestion after remote work
- congestion_index: Combined congestion index

2. Remote Work Dataset (National Remote Work Surveys): Captures the share of remote workers and commuting time saved.
- city: City name
- year: Year of survey
- remote_work_share: Proportion of remote workers
- avg_commute_time_saved: Average daily commute time saved (minutes)

The datasets were merged on city and year for comparative analysis.

2. Challenges Faced

- Missing values in traffic readings due to incomplete records.
- Duplicate rows for overlapping city-year entries.
- Inconsistent city identifiers and year formats.
- Alignment issues between datasets with different time ranges.

3. Data Cleaning Steps

All cleaning operations were performed using pandas in Python.
1. Imported datasets using pd.read_csv().
2. Replaced missing values with mean for traffic data and zero for remote work data.
3. Removed duplicates using drop_duplicates().
4. Standardized city identifiers and year formats for merging.

4. Data Transformation Steps

Several derived variables were created:

1. Traffic Reduction Percentage = ((pre_remote_congestion - post_remote_congestion)

/ pre_remote_congestion) * 100
2. Productivity Ratio = remote_work_share × avg_commute_time_saved
3. Comparison Index = Normalized value of traffic reduction percentage (0–1 scale)
4. Datasets merged on city and year using inner join.
5. Final dataset saved as cleaned_dataset.csv.

5. Output Summary

The final cleaned dataset (cleaned_dataset.csv) includes:
city, year, traffic_reduction_percent, productivity_ratio, and comparison_index.

city	year	traffic_reduction_percent	productivity_ratio	comparison_index
London	2021	12.45	23.6	0.74
New York	2021	9.80	20.2	0.59

6. Key Takeaways

- Data cleaning and transformation ensure reliable analysis for later phases.
- Missing data imputation and duplicate removal improved consistency.
- Derived variables enable insights into the link between remote work and traffic reduction.
- The cleaned dataset will be used in Phase 2 (EDA & Visualization).