Phase 1 Documentation – Group 20

Project Title: Remote Work and Urban Traffic Reduction

Course: Introduction of Data Science

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 investigates how **remote work adoption** impacts **urban traffic congestion** in major global cities.
Two datasets were utilized — one fetched live via **TomTom Traffic API** and the other from **OECD Excel files** related to remote work trends.

(a) Traffic Dataset – TomTom API (Live Data)

Fetched real-time city-level traffic congestion metrics for six major cities.

Attributes:

Column	Description
city	Name of the city
currentSpeed	Current observed speed (km/h)
freeFlowSpeed	Speed during free-flow conditions (no congestion)
confidence	Accuracy/confidence level of data
timestamp	Exact time when data was fetched

Cities Covered: Mumbai, Delhi, Singapore, Dubai, Kuala Lumpur, Riyadh

(b) Remote Work Dataset – OECD Excel Files

These files contained country-level statistics related to remote work adoption. Multiple Excel sheets were merged successfully.

Attributes:

Column	Description
Year	Year of data collection
Indicator	Remote work or productivity metric
ISO	Country code
Country	Country name
Breakdown	Demographic/work type division
Unit of measure	Unit of measurement (percentage/share)
remote_work_share	Share (%) of remote workers in the workforce

Merged Data Shape: (120 rows × 7 columns)

2. Challenges Faced

During the cleaning and merging process, several issues were encountered:

- Missing or null traffic readings from the API for certain timestamps.
- Duplicate rows after merging multiple OECD Excel files.
- Inconsistent column names between OECD datasets.
- Mismatch in city-level vs country-level granularity during merging.
- Deprecated pandas warning (`errors='ignore'`) requiring future syntax updates.

3. Data Cleaning Steps

All data cleaning operations were performed in **Python (pandas)** within the script `scripts/codefile.py`.

Step	Operation	Description
1	API Data Fetching	Fetched live traffic data using TomTom API for six cities.
2	CSV/Excel Import	Merged multiple OECD Excel files using pandas.
3	Missing Value Handling	Replaced missing numeric values with mean (traffic) or zero (remote work).
4	Duplicate Removal	Used `drop_duplicates()` for OECD merged dataset.
5	Column Standardization	Renamed inconsistent headers for merging and analysis.
6	Data Type Correction	Used `pd.to_numeric()` for numerical columns.
7	Final Saving	Saved cleaned traffic and remote work datasets into CSV files.

4. Data Transformation Steps

After cleaning, new derived variables and combined datasets were created for analysis.

Variable	Formula	Description
Traffic Reduction Percentage	((freeFlowSpeed - currentSpeed) / freeFlowSpeed) × 100	Measures congestion level (higher = more congestion reduction potential).
Productivity Ratio	remote_work_share × average commute time saved	Shows estimated productivity benefit of remote work.
Comparison Index	Normalized 0–1 score	Used for comparing both datasets on the same scale.
Merge Logic	city/year	Combined traffic and remote work data conceptually on similar time frames.

5. Output Summary

✅ Both datasets were cleaned, merged, and transformed successfully.
✅ Final metrics were saved automatically in the file:
`/scripts/output_phase1/final_metrics_20251021_074111.csv`

Sample Output:

year	avg_remote_work_share	avg_speed
2018	14.05	48.33

Calculated Metrics (from script logs):

- 🚦 Traffic Reduction Percentage: 4.92%
- 💼 Productivity Ratio: 16.87
- 📈 Comparison Index (0–1 scale): 0.05
