DSA210 Project Presentation

Name: Alp Bartu
Surname: Utar

1. Introduction and Motivation

The goal of this study is to understand how weather (temperature and precipitation) and time-of-day affect hourly coffee-shop customer traffic. By combining cleaned point-of-sale data with local weather records for January 2025, we aim to:

- Quantify the relationship between temperature, rain, and check counts.
- Identify daily and hourly patterns in customer visits.
- Provide actionable insights for staffing and promotions.

2. Data Description

Traffic Data (`cleaned_coffee_shop_data.csv`):

Using raw data t“simpra-saatlik-yogunluk-raporu-01-01-2025-31-01-2025-1739387506.xlsx” after cleaning ,stripping and renaming the columns of the csv file .

- **Columns**: `hour` (0–23), `check_count` (number of customer checks per hour).
- **Rows**: 20 aggregated hourly observations (no dates).

Weather Data (`weather_df.csv`):

- **Columns**: `date` (YYYY-MM-DD), `hour` (0–23), `temp` (°C), `precip` (0/1).
- **Range**: 2025-01-01 to 2025-01-31 (744 hourly records).

3. Methodology

Given the lack of date information in the traffic file, we used the **aggregated per-hour** pipeline:

1. **Merge** average weather by hour (mean `temp`, mean `precip`) with the 20 hourly `check_count` values.
2. **Ordinary Least Squares (OLS) Regression**:

4. Results

4.1 Regression (Aggregated per-hour)

```
OLS Regression Results
Dep. Variable:    check_count   R-squared: 0.531
Model:            OLS            Adj. R-squared: 0.475
----------------------------------------------------------------------------
---
const        70.8265 (p=0.437)
temp_mean    32.1302 (p=0.002)   ← positive and significant
precip_mean2462.1834 (p<0.001)  ← strong effect of rain
```

- **Interpretation**:
  - A 1 °C increase in temperature is associated with ~32 additional checks per hour ($p=0.002$).
  - Rainy hours see ~2462 more checks ($p<0.001$) — likely an artifact of rain coinciding with peak business hours in our small sample.

4.2 Hourly Profile

Bar chart shows peak traffic around **14:00–15:00** and a trough at **03:00–06:00**, consistent with typical business hours.

(In the full date×hour pipeline, we would see):

- **Weekday-hour heatmap** revealing higher weekday midday traffic.
- **Correlations** between weather and traffic, varying by hour.

4.3 Machine Learning Methods on Dataset

1. Overview

I trained two supervised regression models—Linear Regression and Random Forest—on the merged coffee-shop traffic + weather dataset to predict hourly check counts.

---

2. Data & Preprocessing

- **Dataset**: `df_full` with 744 hourly rows (Jan 1–31, 2025), featuring
    - **Features**:
        - `hour` → one-hot dummies (23 columns after drop-first)
        - `temp` (°C), `precip` (0/1)
    - **Target**: `check_count` (integer)
- **Train/Test Split**:
    - **Training**: 595 samples (80%)
    - **Test**: 149 samples (20%)

---

3. Modeling Approach

1. **Feature Encoding**: One-hot encode `hour` (drop the first category to avoid collinearity).
2. **Models**:
    - **LinearRegression** (ordinary least squares)
    - **RandomForestRegressor** (100 trees)
3. **Validation**: 5-fold cross-validation on the training set for R^2 assessment.
4. **Evaluation**: R^2 and RMSE on the hold-out test set.

---

4. Performance Metrics

Model	CV R^2 Mean	CV R^2 Std	Test R^2	Test RMSE
LinearRegression	1.000	0.000	1.000	9.66×10^{-13}
RandomForestRegressor	0.999898	0.000174	0.999997	0.4767

- **Linear Regression** achieves a mathematically "perfect" fit (R^2=1, RMSE≈0), indicating the model has memorized the data.
- **Random Forest** is almost as perfect (Test R^2≈0.999997, RMSE≈0.48), with minimal variance across CV folds.

---

5. Feature Importance (Random Forest Top 5)

Feature	Importance Score
hour_20	0.1080
hour_21	0.0875
hour_15	0.0749
hour_16	0.0744
hour_14	0.0731

The highest-weight features correspond to afternoon/evening hours—peak business periods in January.

---

6. Interpretation & Implications

- **Time-of-Day Dominance**
  The dummy variables for `hour` explain virtually all variance in `check_count`. Weather features (`temp`, `precip`) contribute almost nothing once the model "knows" the hour.
- **Overfitting Warning**
  The near-perfect scores reflect that the model is simply reproducing the known hourly pattern, not uncovering subtle weather effects.
- **Actionable Insight**
  While the model nails your hourly rhythm, it doesn't meaningfully quantify how temperature or rain shifts traffic.

5. Discussion

- **Temperature**: Moderate significant effect; warmer temperatures encourage footfall.
- **Precipitation**: Positive coefficient likely conflates rain with rush periods (requires more granular data).
- **Limitations**:
  - Traffic data lacks date stamps; synthetic replication can distort temporal inference.
  - Small sample (20 hours) limits generalizability.
  - Potential multicollinearity if full dummies used—mitigated by dropping one dummy.