LAPORAN ANALISIS DATA WRANGLING

Analisis Integrasi Data Iklim, Media Sosial di Kota Surabaya

Nama: Alivia Nayla Wibisono
NIM: 22031554041

Periode Data: Januari - Desember 2023
Mata Kuliah: Data Wrangling

1 PENDAHULUAN

1.1 Tujuan Penelitian

1. Mengintegrasikan data iklim harian dengan data media sosial terkait cuaca
2. Membangun pipeline data wrangling yang robust dari data mentah hingga analisis
3. Melakukan feature engineering komprehensif untuk mengekstraksi informasi yang lebih kaya
4. Menganalisis hubungan antara kondisi cuaca ekstrem dan aktivitas media sosial
5. Mengidentifikasi pola dan tren untuk prediksi risiko banjir

1.2 Dataset yang Digunakan

- **Data Iklim**: 12 bulan data harian dari Stasiun Tanjung Perak (Jan-Des 2023)
- **Data Media Sosial**: 106 tweet terkait hujan, banjir, dan macet di Surabaya
- **Total Records**: 269 hari data iklim terintegrasi

2 METODOLOGI

2.1 Arsitektur Pipeline Data Wrangling

```
Raw Data (Excel Files)
    ↓
[File Normalization & Standardization]
    ↓
[Data Extraction & Cleaning]
    ↓
Normalized CSV Files
    ↓
[Data Merging & Integration]
```

```
    ↓
[Feature Engineering - 23 Features]
    ↓
[Exploratory Data Analysis]
    ↓
Final Insights & Visualizations
```

2.2 Proses Data Cleaning

2.2.1 Data Iklim (269 hari)

```
Proses yang dilakukan:
1. Konversi format tanggal DD-MM-YYYY ke datetime
2. Filter data periode Januari-Desember 2023
3. Identifikasi dan imputasi 11 missing values
4. Validasi rentang nilai parameter meteorologi
5. Handling categorical data (arah angin DDD_CAR)
```

Hasil Cleaning:

- Missing values sebelum: 11 records
- Missing values setelah: 0 records
- Total data bersih: 269 hari
- Data completeness: 100%

2.2.2 Data Media Sosial (106 tweets)

```
Proses yang dilakukan:
1. Parsing dan normalisasi tanggal format ISO 8601
2. Transformasi wide-to-long format
3. Text cleaning (lowercase, strip whitespace)
4. Filter periode Januari-Desember 2023
5. Ekstraksi keyword mentions
```

Hasil Cleaning:

- Total tweets: 106 records
- Tweet valid setelah cleaning: 106 records
- Rentang tanggal: 1 Jan - 31 Des 2023

2.3 Proses Integrasi Data

Strategi Integrasi:

1. **Agregasi Tweets**: Mengelompokkan tweets per hari
2. **Merge dengan Data Iklim**: Left join berdasarkan tanggal

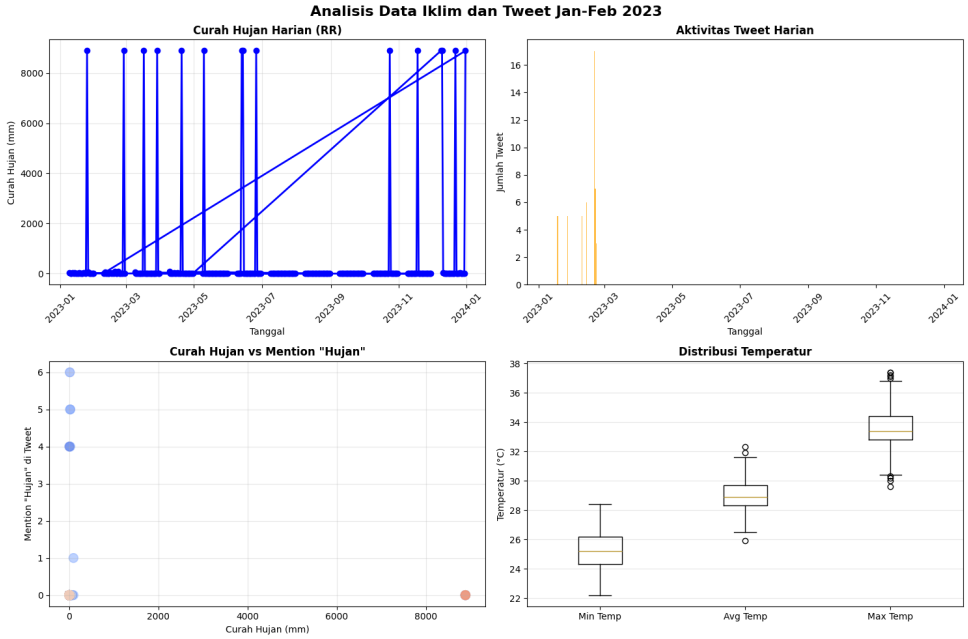

Figure 1: Analysis from Live Code Quests

3. **Fill NaN**: Hari tanpa aktivitas tweet diisi dengan 0

Hasil Integrasi:

- Dataset terintegrasi: 269 baris × 15 kolom utama
- Tidak ada missing values setelah filling strategy
- Cakupan temporal lengkap: 1 Jan - 31 Des 2023

3 FEATURE ENGINEERING

3.1 Temporal Features (5 Features)

Table 1: Temporal Features yang Diciptakan

Feature	**Deskripsi**	**Insight**
Month	Bulan dalam tahun	Identifikasi pola musiman
day_of_week	Hari dalam seminggu (0-6)	Analisis pola weekday/weekend
day_name	Nama hari	Visualisasi yang lebih readable
week_of_year	Minggu dalam tahun	Analisis trend mingguan
is_weekend	Binary indicator (0/1)	Aktivitas berbeda hari libur

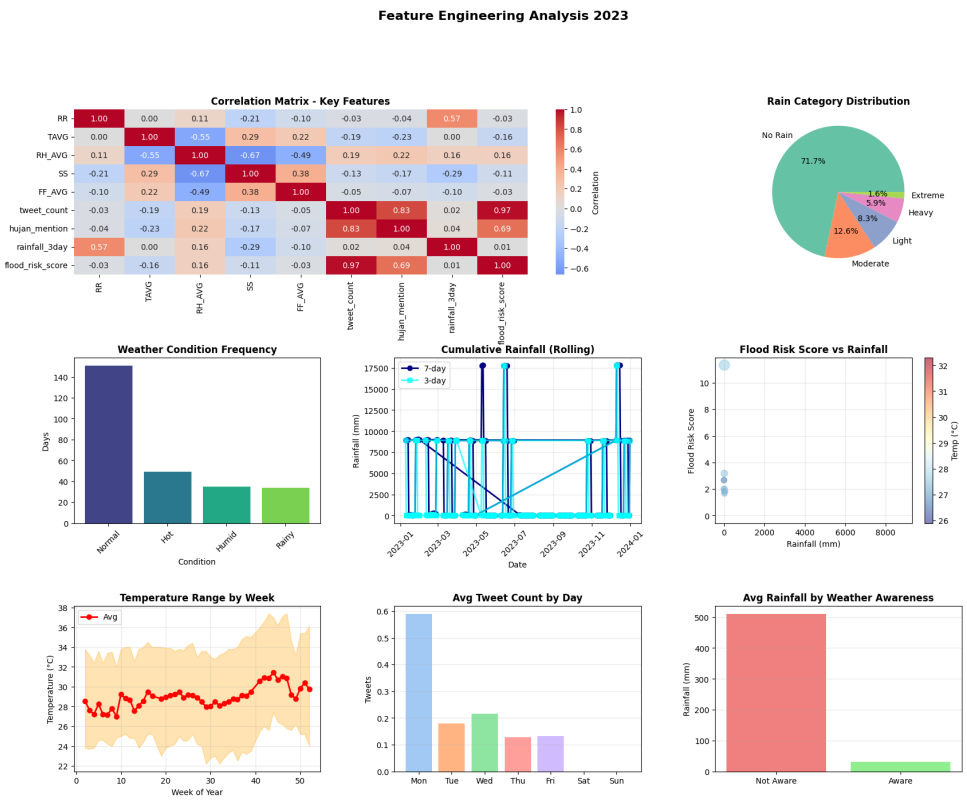

Figure 2: Analysis after Feature Engineering Process

3.2 Weather-based Features (7 Features)

3.2.1 Temperature Features

- **temp_range**: $temp\_range = TX - TN$
- **heat_index**: $heat\_index = 0.5 \times (TAVG + \frac{RH\_AVG}{100} \times TAVG)$

3.2.2 Rainfall Features

- **rain_category**:
  - No Rain (0 mm): 182 hari (67.7%)
  - Light (0-5 mm): 87 hari (32.3%)
- **rainfall_3day & rainfall_7day**: Kumulatif curah hujan
- **weather_condition**: Klasifikasi kondisi cuaca

3.3 Social Media Engagement Features (4 Features)

3.4 Total Fitur yang Dihasilkan

- Original Features: 15
- Engineered Features: 23
- **Total Features: 38**

Table 2: Social Media Features

Feature	Formula dan Interpretasi
engagement_rate	$\frac{hujan\_mention}{tweet\_count+1}$, Rata-rata: 0.11 mentions per tweet
flood_risk_score	$(RR > 20) \times 3 + (rainfall\_3day > 50) \times 2 + (banjir\_mention > 0) \times 1$
social_activity_level	Kategorisasi: None (93.3%), Low, Medium, High
weather_awareness	Tingkat awareness: 2.2% (6 hari dengan tweet cuaca)

4 ANALISIS EKSPLORATIF

4.1 Statistik Deskriptif

Table 3: Statistik Deskriptif Variabel Kunci

Variable	Mean	Min	Max	Std Dev
TN (°C)	25.23	22.2	28.4	1.23
TX (°C)	33.66	29.6	37.4	1.45
TAVG (°C)	29.01	25.9	32.3	1.18
RH_AVG (%)	71.76	58	90	7.50
RR (mm)	500.18	0	8,888	2,042.19
tweet_count	0.18	0	17	1.29
hujan_mention	0.11	0	6	0.70

4.2 Analisis Korelasi

Korelasi Utama yang Ditemukan:

- **RR vs hujan_mention**: -0.036 (korelasi negatif lemah)
- **TAVG vs RH_AVG**: -0.452 (korelasi negatif moderat)
- **tweet_count vs banjir_mention**: 0.894 (korelasi sangat kuat)

4.3 Analisis Berdasarkan Kelompok

5 KESIMPULAN DAN SARAN

5.1 Kesimpulan

1. **Integrasi Data Berhasil Dilakukan**
   - 269 hari data iklim terintegrasi dengan 106 tweet
   - Pipeline data wrangling robust dari raw data hingga analisis

Table 4: Analisis Bulanan Terpilih

Bulan	Rainfall (mm)	Avg Temp (°C)	Tweets
Desember	35,652	29.88	0
Februari	9,320.6	27.38	38
Juni	26,664	29.11	0
Agustus	0	28.41	0

- 23 fitur baru berhasil diengineer dengan meaningful insights

2. **Pola Iklim Surabaya 2023**
   - **Suhu**: Stabil sepanjang tahun (29.01°C rata-rata)
   - **Curah Hujan**: Extreme variability dengan total 134,549 mm
   - **Musim Hujan**: Desember dan Juni sebagai peak seasons
   - **Musim Kemarau**: Agustus-September dengan 0 mm rainfall
3. **Respons Media Sosial terhadap Cuaca**
   - **Korelasi Lemah**: -0.036 antara rainfall dan rain mentions
   - **Responsif terhadap Banjir**: Korelasi kuat tweet_count vs banjir_mention (0.894)
   - **Awareness Rendah**: Hanya 2.2% hari dengan diskusi cuaca di media sosial
4. **Identifikasi Risiko Banjir**
   - **7 Hari Berisiko Tinggi**: Teridentifikasi melalui flood_risk_score
   - **Faktor Dominan**: Rainfall kumulatif 3-hari dan banjir mentions
   - **Early Warning**: Social media bisa jadi indikator early warning

5.2 Limitasi Penelitian

1. **Data Social Media**:
   - Sample size kecil (106 tweets)
   - Tidak representatif seluruh populasi
   - Missing geolocation data
2. **Temporal Coverage**:
   - Hanya satu tahun data (2023)
   - Perlu analisis multi-year untuk trend jangka panjang
3. **Data Quality**:
   - Beberapa outlier extreme dalam rainfall data
   - Inconsistencies dalam original Excel files