preprocess

April 3, 2025

1 Prerocessing Data

[5]:
```
##############################################################################
# data_preprocessing.py
#
# Script that:
#   1) Loads 5 CSV datasets (KSI, TMC, Police Collisions, Env Canada Weather,␣
 ↪ERA5)
#   2) Displays random samples (optional)
#   3) Filters records to [2015-2020]
#   4) Drops unneeded columns (keeps only your specified columns)
#   5) Standardizes & cleans coordinates
#   6) Normalizes temporal data (datetime conversions, timezone removal,
#      combining date + hour in KSI)
#
# NOTE: Removed the __name__ == '__main__' block so this can run from top to␣
 ↪bottom.
##############################################################################

import pandas as pd
import numpy as np
import ast
from IPython.display import display, Markdown
from datetime import datetime

##############################################################################
# 1) LOAD DATA
##############################################################################

def load_datasets():
    print("Loading datasets...")
    df_ksi = pd.read_csv("Motor Vehicle Collisions with KSI Data - 4326.csv")
    df_tmc = pd.read_csv("tmc_raw_data_2010_2019.csv")
    df_collisions = pd.read_csv("Traffic_Collisions_Toronto_data.csv")
    df_env = pd.read_csv("hourly_final.csv")  # Env Canada
    df_era5 = pd.read_csv("ERA.csv")          # ERA5 from GEE
    return df_ksi, df_tmc, df_collisions, df_env, df_era5
```

```
##############################################################################
# 2) SHOW RANDOM UNIQUE ROWS (Optional Inspection)
##############################################################################

def show_random_unique_rows(df, title, n=10, subset=None):
    """
    Displays n unique random rows from df. If 'subset' is given,
    ensures uniqueness by those columns first, then samples.
    """
    if subset:
        df = df.drop_duplicates(subset=subset)
    else:
        df = df.drop_duplicates()
    sample = df.sample(min(n, len(df)), random_state=42)
    display(Markdown(f"### {title} - {len(sample)} Unique Rows"))
    display(sample.reset_index(drop=True))

def inspect_data(df_ksi, df_tmc, df_collisions, df_env, df_era5):
    """Display random samples from each dataset for a quick look."""
    show_random_unique_rows(df_ksi, "KSI Dataset", n=10,␣
 ↪subset=["STREET1","STREET2"])
    show_random_unique_rows(df_tmc, "TMC Dataset", n=10,␣
 ↪subset=["location_name"])
    show_random_unique_rows(df_collisions, "Collisions Dataset", n=10,␣
 ↪subset=["Neighbourhood"])
    show_random_unique_rows(df_env, "Env Canada Dataset", n=10)
    show_random_unique_rows(df_era5, "ERA5 Dataset", n=10)

##############################################################################
# 3) DATE/TIME FILTER (2015-2020)
##############################################################################

def to_datetime_and_filter(df, date_col, start="2015-01-01", end="2020-12-31"):
    """
    Convert df[date_col] to datetime, remove timezone, filter rows to [start,␣
 ↪end].
    """
    df[date_col] = pd.to_datetime(df[date_col], errors='coerce')
    df[date_col] = df[date_col].dt.tz_localize(None)  # remove tz if present
    mask = (df[date_col] >= pd.Timestamp(start)) & (df[date_col] <= pd.
 ↪Timestamp(end))
    return df[mask].copy()

##############################################################################
# 4) DROP COLUMNS (KEEP LISTS)
##############################################################################
```

```
ksi_final_keep_cols = [
    'DATE', 'TIME', 'geometry',
    'ACCLASS', 'INJURY', 'FATAL_NO',
    'IMPACTYPE', 'ROAD_CLASS', 'ACCLOC',
    'TRAFFCTL', 'VISIBILITY', 'LIGHT', 'RDSFCOND',
    'VEHTYPE', 'INVTYPE',
    'AUTOMOBILE', 'PEDESTRIAN', 'CYCLIST',
    'NEIGHBOURHOOD_158'
]

tmc_keep_cols = [
    'count_date', 'start_time', 'end_time',
    'longitude', 'latitude',
    'location_name'
]

collision_keep_cols = [
    'OccurrenceDate', 'Hour',
    'Longitude', 'Latitude', 'Neighbourhood',
    'Fatalities', 'Injury_Collisions', 'PD_Collisions'
]

env_keep_cols = [
    'LOCAL_DATE', 'LOCAL_HOUR', 'TEMP', 'WINDCHILL',
    'PRECIP_AMOUNT', 'RELATIVE_HUMIDITY', 'VISIBILITY',
    'WEATHER_ENG_DESC', 'x', 'y'
]

era5_keep_cols = [
    'timestamp', 'temperature_2m', 'dewpoint_temperature_2m',
    'total_precipitation', 'u_component_of_wind_10m',
    'v_component_of_wind_10m', '.geo'
]

def drop_unneeded_columns(df_ksi, df_tmc, df_collisions, df_env, df_era5):
    """
    Drop all columns except those in keep-lists.
    """
    # KSI
    df_ksi = df_ksi[[c for c in ksi_final_keep_cols if c in df_ksi.columns]].
 ↪copy()

    # TMC
    df_tmc = df_tmc[[c for c in tmc_keep_cols if c in df_tmc.columns]].copy()

    # Collisions
```

```
    df_collisions = df_collisions[[c for c in collision_keep_cols if c in␣
 ↪df_collisions.columns]].copy()

    # Env Canada
    df_env = df_env[[c for c in env_keep_cols if c in df_env.columns]].copy()

    # ERA5
    df_era5 = df_era5[[c for c in era5_keep_cols if c in df_era5.columns]].
 ↪copy()

    return df_ksi, df_tmc, df_collisions, df_env, df_era5

################################################################################
# 5) STANDARDIZE & CLEAN COORDINATES
################################################################################

def parse_ksi_geometry(geom_str):
    """Parses KSI 'geometry' JSON to extract (lon, lat)."""
    if pd.isna(geom_str):
        return pd.Series([None, None])
    try:
        geom_dict = ast.literal_eval(geom_str)
        coords = geom_dict.get("coordinates", None)
        if isinstance(coords, list) and len(coords) > 0:
            first_elem = coords[0]
            if isinstance(first_elem, list) and len(first_elem) > 0 and␣
 ↪isinstance(first_elem[0], list):
                # Possibly a MultiLineString
                first_pair = first_elem[0]
            else:
                first_pair = first_elem
            if isinstance(first_pair, list) and len(first_pair) == 2:
                return pd.Series([float(first_pair[0]), float(first_pair[1])])
    except:
        pass
    return pd.Series([None, None])

def standardize_coordinates(df_ksi, df_tmc, df_collisions, df_env, df_era5):
    """
    For each dataset:
      - KSI: parse geometry -> (lon, lat)
      - TMC: rename (longitude->lon, latitude->lat)
      - Collisions: rename (Longitude->lon, Latitude->lat)
      - Env Canada: rename (x->lon, y->lat)
      - ERA5: parse .geo -> (lon, lat)
    """
    # --- KSI ---
```

```
    if 'geometry' in df_ksi.columns:
        df_ksi[['lon','lat']] = df_ksi['geometry'].apply(parse_ksi_geometry)
        df_ksi.drop(columns=['geometry'], inplace=True, errors='ignore')

    # --- TMC ---
    if 'longitude' in df_tmc.columns:
        df_tmc.rename(columns={'longitude':'lon','latitude':'lat'},␣
 ↪inplace=True)

    # --- Collisions ---
    if 'Longitude' in df_collisions.columns:
        df_collisions.rename(columns={'Longitude':'lon'}, inplace=True)
    if 'Latitude' in df_collisions.columns:
        df_collisions.rename(columns={'Latitude':'lat'}, inplace=True)

    # --- Env Canada ---
    if 'x' in df_env.columns and 'y' in df_env.columns:
        df_env.rename(columns={'x':'lon','y':'lat'}, inplace=True)

    # --- ERA5 ---
    if '.geo' in df_era5.columns:
        def parse_era5_geo(geo_str):
            if pd.isna(geo_str):
                return pd.Series([None, None])
            try:
                geo_dict = ast.literal_eval(geo_str)
                coords = geo_dict.get("coordinates", None)
                if coords and len(coords) == 2:
                    return pd.Series([float(coords[0]), float(coords[1])])
            except:
                pass
            return pd.Series([None, None])

        df_era5[['lon','lat']] = df_era5['.geo'].apply(parse_era5_geo)
        df_era5.drop(columns=['.geo'], inplace=True)

    return df_ksi, df_tmc, df_collisions, df_env, df_era5

################################################################################
# 6) NORMALIZE TEMPORAL DATA (Datetime)
################################################################################

def normalize_temporal_data(df_ksi, df_tmc, df_collisions, df_env, df_era5):
    """
    For each dataset:
      - Convert date/time fields to datetime
      - Combine KSI (DATE + TIME) into single 'datetime'
```

```
    - Round or remove tz if needed
    """
    # --- KSI ---
    if 'DATE' in df_ksi.columns and 'TIME' in df_ksi.columns:
        def combine_ksi_time(row):
            if pd.isna(row['DATE']):
                return pd.NaT
            t_str = str(row['TIME']).zfill(4)
            hh = int(t_str[:-2]) if len(t_str) >= 2 else 0
            mm = int(t_str[-2:]) if len(t_str) >= 2 else 0
            return row['DATE'] + pd.Timedelta(hours=hh, minutes=mm)

        df_ksi['datetime'] = df_ksi.apply(combine_ksi_time, axis=1)
        df_ksi['datetime'] = pd.to_datetime(df_ksi['datetime'], errors='coerce')
        df_ksi['datetime'] = df_ksi['datetime'].dt.round('H')  # rounding to␣
↪nearest hour

    # --- TMC ---
    if 'start_time' in df_tmc.columns:
        df_tmc['start_time'] = pd.to_datetime(df_tmc['start_time'],␣
↪errors='coerce')
        df_tmc['start_time'] = df_tmc['start_time'].dt.tz_localize(None).dt.
↪round('H')

    if 'count_date' in df_tmc.columns:
        df_tmc['count_date'] = pd.to_datetime(df_tmc['count_date'],␣
↪errors='coerce')
        df_tmc['count_date'] = df_tmc['count_date'].dt.tz_localize(None)

    # --- Collisions ---
    if 'OccurrenceDate' in df_collisions.columns:
        df_collisions['OccurrenceDate'] = pd.
↪to_datetime(df_collisions['OccurrenceDate'], errors='coerce')
        df_collisions['OccurrenceDate'] = df_collisions['OccurrenceDate'].dt.
↪tz_localize(None).dt.round('H')

    # --- Env Canada ---
    if 'LOCAL_DATE' in df_env.columns:
        df_env['LOCAL_DATE'] = pd.to_datetime(df_env['LOCAL_DATE'],␣
↪errors='coerce')
        df_env['LOCAL_DATE'] = df_env['LOCAL_DATE'].dt.tz_localize(None).dt.
↪round('H')

    # --- ERA5 ---
    if 'timestamp' in df_era5.columns:
```

```
        df_era5['timestamp'] = pd.to_datetime(df_era5['timestamp'],␣
 ↪errors='coerce')
        df_era5['timestamp'] = df_era5['timestamp'].dt.tz_localize(None).dt.
 ↪round('H')

    return df_ksi, df_tmc, df_collisions, df_env, df_era5

################################################################################
# EXECUTION (TOP-TO-BOTTOM SCRIPT)
################################################################################

# STEP 1) Load Data
df_ksi, df_tmc, df_collisions, df_env, df_era5 = load_datasets()

# STEP 2) Inspect (Optional)
inspect_data(df_ksi, df_tmc, df_collisions, df_env, df_era5)

# STEP 3) Filter to 2015-2020
df_ksi = to_datetime_and_filter(df_ksi, 'DATE')
df_tmc = to_datetime_and_filter(df_tmc, 'count_date')
df_collisions = to_datetime_and_filter(df_collisions, 'OccurrenceDate')
df_env = to_datetime_and_filter(df_env, 'LOCAL_DATE')
df_era5 = to_datetime_and_filter(df_era5, 'timestamp')

# STEP 4) Drop Unneeded Columns
df_ksi, df_tmc, df_collisions, df_env, df_era5 = drop_unneeded_columns(
    df_ksi, df_tmc, df_collisions, df_env, df_era5
)

# STEP 5) Standardize & Clean Coordinates
df_ksi, df_tmc, df_collisions, df_env, df_era5 = standardize_coordinates(
    df_ksi, df_tmc, df_collisions, df_env, df_era5
)

# STEP 6) Normalize Temporal Data
df_ksi, df_tmc, df_collisions, df_env, df_era5 = normalize_temporal_data(
    df_ksi, df_tmc, df_collisions, df_env, df_era5
)

# PRINT final shapes and heads
print("\n===== FINAL DATAFRAMES =====")
print(f"KSI shape: {df_ksi.shape}")
print(f"TMC shape: {df_tmc.shape}")
print(f"Collisions shape: {df_collisions.shape}")
print(f"Env Canada shape: {df_env.shape}")
print(f"ERA5 shape: {df_era5.shape}")
```

```
print("\n--- KSI HEAD ---")
print(df_ksi.head())

print("\n--- TMC HEAD ---")
print(df_tmc.head())

print("\n--- Collisions HEAD ---")
print(df_collisions.head())

print("\n--- Env Canada HEAD ---")
print(df_env.head())

print("\n--- ERA5 HEAD ---")
print(df_era5.head())
```

```
Loading datasets…

C:\Users\manav\AppData\Local\Temp\ipykernel_5196\2960273254.py:31: DtypeWarning:
Columns (4,8,17,19,20,24,34) have mixed types. Specify dtype option on import or
set low_memory=False.
  df_env = pd.read_csv("hourly_final.csv")  # Env Canada
```

1.0.1 KSI Dataset – 10 Unique Rows

```
     _id        ACCNUM        DATE  TIME              STREET1  \
0   4447  1.097712e+06  2009-04-04   228        LIPPINCOTT ST
1    259  8.991260e+05  2006-03-26  1614         CALEDONIA RD
2  11502           NaN  2015-04-22  1402         3 RAINIER SQ
3  15991           NaN  2019-10-05  1611            DUPONT ST
4  15656           NaN  2019-06-20  1315     REGENT PARK BLVD
5  16973  1.000499e+09  2021-03-18  1056  GREENWIN VILLAGE RD
6   6095  1.180034e+06  2010-07-16  1011          SPADINA AVE
7   8566  1.315841e+06  2012-08-12   951            CHURCH ST
8   2426  9.904020e+05  2007-08-27  1441         FINCH Aven E
9   7158  1.251049e+06  2011-07-17   630         BREMNER Boul

            STREET2        OFFSET      ROAD_CLASS                DISTRICT  \
0         BLOOR ST W           NaN  Major Arterial  Toronto and East York
1         NORMAN AVE           NaN  Minor Arterial  Toronto and East York
2                NaN  4 m South of           Local             Scarborough
3         BEDFORD RD           NaN  Minor Arterial  Toronto and East York
4                NaN           NaN  Major Arterial  Toronto and East York
5        BATHURST ST  50 m East of       Collector              North York
6       ST ANDREW ST           NaN  Major Arterial  Toronto and East York
7         CARLTON ST           NaN  Major Arterial  Toronto and East York
8        LESLIE Stre           NaN  Major Arterial              North York
9  VAN DE WATER Cres           NaN  Minor Arterial  Toronto and East York

                ACCLOC  … AG_DRIV REDLIGHT ALCOHOL DISABILITY HOOD_158  \
```

```
0                   NaN  …     NaN      NaN     Yes        NaN       79
1                   NaN  …     NaN      NaN     NaN        NaN       92
2       Private Driveway  …     Yes      NaN     NaN        NaN      148
3        At Intersection  …     NaN      NaN     NaN        NaN       95
4        At Intersection  …     NaN      NaN     NaN        NaN       72
5  At/Near Private Drive  …     Yes      NaN     NaN        NaN       36
6                   NaN  …     NaN      NaN     NaN        NaN       78
7        At Intersection  …     Yes      Yes     NaN        NaN      168
8        At Intersection  …     NaN      NaN     NaN        NaN       49
9        At Intersection  …     Yes      NaN     NaN        NaN      165

       NEIGHBOURHOOD_158 HOOD_140                     NEIGHBOURHOOD_140  \
0             University       79                       University (79)
1  Corso Italia-Davenport       92            Corso Italia-Davenport (92)
2        East L'Amoreaux      117                      L'Amoreaux (117)
3                 Annex       95                             Annex (95)
4            Regent Park       72                       Regent Park (72)
5       Newtonbrook West       36                  Newtonbrook West (36)
6   Kensington-Chinatown       78               Kensington-Chinatown (78)
7    Downtown Yonge East       75              Church-Yonge Corridor (75)
8   Bayview Woods-Steeles       49              Bayview Woods-Steeles (49)
9  Harbourfront-CityPlace       77  Waterfront Communities-The Island (77)

  DIVISION                                           geometry
0      D14  {"coordinates": [[-79.4099900003758, 43.665344…
1      D13  {"coordinates": [[-79.4557899995631, 43.677744…
2      D42  {"coordinates": [[-79.3030119995988, 43.794242…
3      D53  {"coordinates": [[-79.4005079994135, 43.676299…
4      D51  {"coordinates": [[-79.3617430003326, 43.660287…
5      D32  {"coordinates": [[-79.4448350001569, 43.790754…
6      D52  {"coordinates": [[-79.3985930000871, 43.654345…
7      D51  {"coordinates": [[-79.3793900004206, 43.661845…
8      D33  {"coordinates": [[-79.3680900003581, 43.790045…
9      D52  {"coordinates": [[-79.3889900000155, 43.640345…

[10 rows x 50 columns]
```

1.0.2 TMC Dataset – 10 Unique Rows

```
      _id  count_id  count_date  \
0   65665     29519  2012-12-04
1   42623     28013  2011-10-17
2  114332     32687  2015-07-20
3   95360     31459  2014-10-16
4  152610     35121  2016-11-03
5  165171     35931  2017-05-09
6   32162     27291  2011-05-02
7  147353     34773  2016-11-09
```

```
8   12082     25913  2010-05-18
9  110891     32461  2015-05-06

                                      location_name  longitude   latitude  \
0                            King St W / Stanley Ter -79.410122  43.642417
1                          Comstock Rd / Pharmacy Ave -79.294925  43.719318
2  Lake Shore Blvd W / Brown's Line / Thirty Eigh… -79.539402  43.593100
3                          Windermere Ave / Annette St -79.483910  43.659196
4                         Torbarrie Rd / Judy Sgro Ave -79.523647  43.727535
5                           Indian Rd / High Park Blvd -79.453168  43.645499
6   Willowdale Ave / Bishop Ave / Finch Corridor Trl -79.407872  43.783441
7                  Queen's Park Cres E / St Joseph St -79.391135  43.664851
8                           Bloor St W / Lansdowne Ave -79.442734  43.658338
9  Warden Ave / Clonmore Dr / Hollis Kalmar Park Trl -79.273203  43.693282

   centreline_type  centreline_id      px           start_time  …  \
0                2       13467925  2314.0  2012-12-04T07:30:00  …
1                2       13454827   941.0  2011-10-17T07:30:00  …
2                2       13470669     NaN  2015-07-20T07:30:00  …
3                2       13465467     NaN  2014-10-16T07:30:00  …
4                2       20145216     NaN  2016-11-03T07:30:00  …
5                2       13467619     NaN  2017-05-09T07:30:00  …
6                2       13445655  1640.0  2011-05-02T07:30:00  …
7                2       13464371  2318.0  2016-11-09T07:30:00  …
8                2       13465512   326.0  2010-05-18T07:30:00  …
9                2       13459016     NaN  2015-05-06T07:30:00  …

  w_appr_bus_t  w_appr_bus_l  n_appr_peds  s_appr_peds  e_appr_peds  \
0            7             0            0            0            0
1            0             0            0            0            0
2            3             2            5            4            0
3            3             0            1            1            0
4            0             0            0            0            0
5            1             0            2            8            4
6            0             0            3            0            0
7            0             0            2           12           28
8            0             0           36            0           28
9            0             0            2            2            2

   w_appr_peds  n_appr_bike  s_appr_bike  e_appr_bike  w_appr_bike
0            2            0            0            0            9
1            0            0            0            0            0
2            1            0            0            4            8
3            3            0            1            2            6
4            0            0            0            0            0
5            0            3            1            1            6
6            1            0            0            0            2
7            0            0            0            0            0
```

```
8            0           17            0            9            0
9            2            0            0            0            0

[10 rows x 55 columns]
```

1.0.3 Collisions Dataset – 10 Unique Rows

```
              X             Y  OBJECTID   EventUniqueId  \
0 -8.832802e+06  5.416546e+06    211103  GO-20142589039
1 -8.843931e+06  5.417016e+06    371464  GO-20158008179
2  0.000000e+00  0.000000e+00     69459  GO-20191146878
3 -8.828159e+06  5.413933e+06    161859  GO-20208017470
4 -8.840497e+06  5.431424e+06    330701  GO-20158046140
5 -8.856014e+06  5.421364e+06    473327  GO-20148004896
6 -8.837780e+06  5.413070e+06    279912   GO-2020796549
7 -8.846494e+06  5.419266e+06    397302   GO-2021888119
8 -8.854640e+06  5.420960e+06    462993  GO-20168056458
9  0.000000e+00  0.000000e+00     22395   GO-2017417456

           OccurrenceDate     Month Day_of_Week  Year  Hour Division Atom  \
0  2014/07/28 04:00:00+00      July      Monday  2014    14  D54/D55   57
1  2015/02/20 05:00:00+00  February      Friday  2015    11      D13  109
2  2019/06/20 04:00:00+00      June    Thursday  2019    20      D11   88
3  2020/06/09 04:00:00+00      June     Tuesday  2020    13  D54/D55   70
4  2015/11/05 05:00:00+00  November    Thursday  2015    15      D32   36
5  2014/02/01 05:00:00+00  February    Saturday  2014    18      D23    4
6  2020/04/27 04:00:00+00     April      Monday  2020    17      D52   79
7  2021/05/14 04:00:00+00       May      Friday  2021     0      D12   30
8  2016/12/15 05:00:00+00  December    Thursday  2016    14      D23    5
9  2017/03/07 05:00:00+00     March     Tuesday  2017    18      D43  136

              Neighbourhood  Fatalities Injury_Collisions FTR_Collisions  \
0      Broadview North (57)           0               YES             NO
1  Caledonia-Fairbank (109)           0                NO             NO
2      High Park North (88)           0                NO            YES
3      South Riverdale (70)           0                NO             NO
4     Newtonbrook West (36)           0                NO             NO
5       Rexdale-Kipling (4)           0                NO             NO
6           University (79)           0               YES             NO
7  Brookhaven-Amesbury (30)           0                NO            YES
8      Elms-Old Rexdale (5)           0                NO             NO
9           West Hill (136)           0               YES             NO

  PD_Collisions  Longitude   Latitude  ObjectId2
0            NO -79.346409  43.683194     211120
1           YES -79.446384  43.686249     371222
2            NO   0.000000   0.000000      69871
3           YES -79.304700  43.666220     161631
```

4	YES	-79.415536	43.779769	330003
5	YES	-79.554925	43.714484	473037
6	NO	-79.391125	43.660609	279055
7	YES	-79.469409	43.700861	397039
8	YES	-79.542589	43.711865	462430
9	NO	0.000000	0.000000	22501

1.0.4 Env Canada Dataset – 10 Unique Rows

	x	y	LOCAL_DATE	STATION_PRESSURE	TEMP_FLAG
0	-79.4	43.666667	2018-11-23 17:00:00	100.79	NaN
1	-79.4	43.666667	2024-06-04 15:00:00	99.96	NaN
2	-79.4	43.666667	2020-06-24 16:00:00	99.37	NaN
3	-79.4	43.666667	2018-02-15 13:00:00	99.11	NaN
4	-79.4	43.666667	2014-05-12 17:00:00	100.44	NaN
5	-79.4	43.666667	2017-01-14 08:00:00	102.25	NaN
6	-79.4	43.666667	2025-02-15 22:00:00	99.70	NaN
7	-79.4	43.666667	2017-03-28 12:00:00	100.16	NaN
8	-79.4	43.666667	2019-07-28 18:00:00	100.07	NaN
9	-79.4	43.666667	2015-02-08 19:00:00	100.12	NaN

	WINDCHILL	LOCAL_HOUR	RELATIVE_HUMIDITY	WIND_DIRECTION_FLAG
0	NaN	17	60.0	M
1	NaN	15	56.0	NaN
2	NaN	16	34.0	NaN
3	NaN	13	75.0	M
4	NaN	17	37.0	M
5	NaN	8	66.0	M
6	NaN	22	88.0	NaN
7	NaN	12	69.0	M
8	NaN	18	49.0	NaN
9	NaN	19	84.0	M

	WIND_DIRECTION	…	LOCAL_DAY	PROVINCE_CODE	UTC_DATE
0	NaN	…	23	ON	2018-11-23T22:00:00
1	NaN	…	4	ON	2024-06-04T20:00:00
2	NaN	…	24	ON	2020-06-24T21:00:00
3	NaN	…	15	ON	2018-02-15T18:00:00
4	NaN	…	12	ON	2014-05-12T22:00:00
5	NaN	…	14	ON	2017-01-14T13:00:00
6	NaN	…	15	ON	2025-02-16T03:00:00
7	NaN	…	28	ON	2017-03-28T17:00:00
8	NaN	…	28	ON	2019-07-28T23:00:00
9	NaN	…	8	ON	2015-02-09T00:00:00

	DEW_POINT_TEMP	TEMP	WINDCHILL_FLAG	VISIBILITY	RELATIVE_HUMIDITY_FLAG
0	-6.0	0.8	NaN	NaN	NaN
1	16.9	26.4	NaN	NaN	NaN

```
2             6.2  22.7           NaN         NaN                    NaN
3             3.8   7.9           NaN         NaN                    NaN
4             2.0  16.7           NaN         NaN                    NaN
5           -13.2  -7.9           NaN         NaN                    NaN
6            -4.4  -2.7           NaN         NaN                    NaN
7             2.6   8.0           NaN         NaN                    NaN
8            16.0  27.5           NaN         NaN                    NaN
9           -12.9 -10.6           NaN         NaN                    NaN

  HUMIDEX VISIBILITY_FLAG
0     NaN             NaN
1    32.0             NaN
2     NaN             NaN
3     NaN             NaN
4     NaN             NaN
5     NaN             NaN
6     NaN             NaN
7     NaN             NaN
8    32.0             NaN
9     NaN             NaN

[10 rows x 37 columns]
```

1.0.5 ERA5 Dataset – 10 Unique Rows

```
    system:index  dewpoint_temperature_2m    location  surface_pressure  \
0  9_20190303T01               268.240677   Agincourt     100170.890625
1  4_20200706T14               291.267044   East York     100313.261719
2  5_20151230T10               273.556503   Guildwood     100968.562500
3  5_20200510T02               269.013580   Guildwood     100428.437500
4  1_20201202T02               269.730377  North York      98513.074219
5  5_20151017T00               273.650330   Guildwood     100559.398438
6  6_20150111T13               261.082565   High Park     101304.058594
7  3_20170814T08               288.141953   Etobicoke      99507.597656
8  4_20191107T13               270.301544   East York     100867.046875
9  4_20190805T01               289.511612   East York     100103.890625

   temperature_2m         timestamp  total_precipitation  \
0      270.709167  2019-03-03 01:00         2.066698e-05
1      296.455444  2020-07-06 14:00         8.583069e-07
2      275.845764  2015-12-30 10:00         2.257153e-04
3      276.657867  2020-05-10 02:00         0.000000e+00
4      272.274414  2020-12-02 02:00         5.876273e-05
5      281.555420  2015-10-17 00:00         7.847980e-04
6      265.088928  2015-01-11 13:00         4.999340e-06
7      291.820892  2017-08-14 08:00         0.000000e+00
8      273.623306  2019-11-07 13:00         6.001830e-03
9      295.110657  2019-08-05 01:00         0.000000e+00
```

```
   u_component_of_wind_10m  v_component_of_wind_10m  \
0                 2.189865                -3.053253
1                -4.281723                 0.109268
2                 5.781815                 3.189896
3                 6.243011                -1.323898
4                 4.092887                -4.099899
5                 7.381165                -4.987549
6                 3.057251                 3.387320
7                 0.243805                -0.874802
8                -0.419662                -4.553178
9                 0.110672                 0.722921

                                                 .geo
0  {"type":"Point","coordinates":[-79.2939,43.7996]}
1   {"type":"Point","coordinates":[-79.3017,43.665]}
2  {"type":"Point","coordinates":[-79.1845,43.7636]}
3  {"type":"Point","coordinates":[-79.1845,43.7636]}
4  {"type":"Point","coordinates":[-79.5181,43.7731]}
5  {"type":"Point","coordinates":[-79.1845,43.7636]}
6  {"type":"Point","coordinates":[-79.4309,43.6816]}
7  {"type":"Point","coordinates":[-79.62050000000…
8   {"type":"Point","coordinates":[-79.3017,43.665]}
9   {"type":"Point","coordinates":[-79.3017,43.665]}


===== FINAL DATAFRAMES =====
KSI shape: (5571, 21)
TMC shape: (122570, 6)
Collisions shape: (417795, 8)
Env Canada shape: (52355, 10)
ERA5 shape: (525850, 8)

--- KSI HEAD ---
            DATE  TIME           ACCLASS INJURY  FATAL_NO         IMPACTYPE  \
11303 2015-01-01   624  Non-Fatal Injury  Major       NaN       Approaching
11304 2015-01-01   624  Non-Fatal Injury  Minor       NaN       Approaching
11305 2015-01-01   624  Non-Fatal Injury  Minor       NaN       Approaching
11306 2015-01-02   949  Non-Fatal Injury  Major       NaN  Turning Movement
11307 2015-01-02   949  Non-Fatal Injury    NaN       NaN  Turning Movement

           ROAD_CLASS            ACCLOC    TRAFFCTL VISIBILITY  … RDSFCOND  \
11303  Major Arterial   At Intersection  No Control      Clear  …      Dry
11304  Major Arterial   At Intersection  No Control      Clear  …      Dry
11305  Major Arterial   At Intersection  No Control      Clear  …      Dry
11306  Major Arterial  Non Intersection  No Control      Clear  …      Dry
11307  Major Arterial  Non Intersection  No Control      Clear  …      Dry
```

```
                            VEHTYPE       INVTYPE AUTOMOBILE PEDESTRIAN  \
11303     Automobile, Station Wagon        Driver        Yes        NaN
11304     Automobile, Station Wagon        Driver        Yes        NaN
11305                           NaN     Passenger        Yes        NaN
11306     Automobile, Station Wagon        Driver        Yes        NaN
11307  Truck - Closed (Blazer, etc)  Truck Driver        Yes        NaN

      CYCLIST    NEIGHBOURHOOD_158        lon        lat            datetime
11303     NaN  Brookhaven-Amesbury -79.477496  43.706175 2015-01-01 06:00:00
11304     NaN  Brookhaven-Amesbury -79.477496  43.706175 2015-01-01 06:00:00
11305     NaN  Brookhaven-Amesbury -79.477496  43.706175 2015-01-01 06:00:00
11306     NaN        Junction Area -79.475185  43.670848 2015-01-02 10:00:00
11307     NaN        Junction Area -79.475185  43.670848 2015-01-02 10:00:00

[5 rows x 21 columns]

--- TMC HEAD ---
      count_date          start_time             end_time       lon  \
99531 2015-01-06 2015-01-06 08:00:00  2015-01-06T07:45:00 -79.49892
99532 2015-01-06 2015-01-06 08:00:00  2015-01-06T08:00:00 -79.49892
99533 2015-01-06 2015-01-06 08:00:00  2015-01-06T08:15:00 -79.49892
99534 2015-01-06 2015-01-06 08:00:00  2015-01-06T08:30:00 -79.49892
99535 2015-01-06 2015-01-06 08:00:00  2015-01-06T08:45:00 -79.49892

            lat                 location_name
99531  43.617497  Royal York Rd / Newcastle St
99532  43.617497  Royal York Rd / Newcastle St
99533  43.617497  Royal York Rd / Newcastle St
99534  43.617497  Royal York Rd / Newcastle St
99535  43.617497  Royal York Rd / Newcastle St

--- Collisions HEAD ---
         OccurrenceDate  Hour  lon  lat Neighbourhood  Fatalities  \
9000 2015-01-28 05:00:00     0  0.0  0.0           NSA           0
9001 2015-09-16 04:00:00    17  0.0  0.0           NSA           0
9002 2015-01-28 05:00:00     6  0.0  0.0           NSA           1
9003 2015-09-18 04:00:00    11  0.0  0.0           NSA           0
9004 2015-09-18 04:00:00    11  0.0  0.0           NSA           0

     Injury_Collisions PD_Collisions
9000               YES            NO
9001               YES            NO
9002                NO            NO
9003               YES            NO
9004               YES            NO

--- Env Canada HEAD ---
               LOCAL_DATE  LOCAL_HOUR  TEMP  WINDCHILL  PRECIP_AMOUNT  \
```

```
36750 2020-12-31 00:00:00           0   3.2        NaN            0.0
36751 2020-12-30 23:00:00          23   3.5        NaN            0.0
36752 2020-12-30 22:00:00          22   4.1        NaN            0.0
36753 2020-12-30 21:00:00          21   5.3        NaN            0.0
36754 2020-12-30 20:00:00          20   5.6        NaN            0.1

       RELATIVE_HUMIDITY  VISIBILITY  WEATHER_ENG_DESC   lon        lat
36750               63.0         NaN               NaN -79.4  43.666667
36751               66.0         NaN               NaN -79.4  43.666667
36752               70.0         NaN               NaN -79.4  43.666667
36753               79.0         NaN               NaN -79.4  43.666667
36754               86.0         NaN               NaN -79.4  43.666667

--- ERA5 HEAD ---
            timestamp  temperature_2m  dewpoint_temperature_2m  \
0 2015-01-01 00:00:00      267.469437               256.379730
1 2015-01-01 01:00:00      267.390060               256.667465
2 2015-01-01 02:00:00      267.294342               256.784409
3 2015-01-01 03:00:00      267.115387               256.670761
4 2015-01-01 04:00:00      266.850189               257.201431

   total_precipitation  u_component_of_wind_10m  v_component_of_wind_10m  \
0             0.000267                 7.538605                 3.411713
1             0.000007                 7.506500                 3.796295
2             0.000015                 7.861252                 3.563995
3             0.000021                 8.119202                 3.096085
4             0.000023                 7.925232                 2.782425

       lon      lat
0 -79.3832  43.6532
1 -79.3832  43.6532
2 -79.3832  43.6532
3 -79.3832  43.6532
4 -79.3832  43.6532

C:\Users\manav\AppData\Local\Temp\ipykernel_5196\2960273254.py:226:
FutureWarning: 'H' is deprecated and will be removed in a future version, please
use 'h' instead.
  df_ksi['datetime'] = df_ksi['datetime'].dt.round('H')  # rounding to nearest
hour
C:\Users\manav\AppData\Local\Temp\ipykernel_5196\2960273254.py:231:
FutureWarning: 'H' is deprecated and will be removed in a future version, please
use 'h' instead.
  df_tmc['start_time'] = df_tmc['start_time'].dt.tz_localize(None).dt.round('H')
C:\Users\manav\AppData\Local\Temp\ipykernel_5196\2960273254.py:240:
FutureWarning: 'H' is deprecated and will be removed in a future version, please
use 'h' instead.
  df_collisions['OccurrenceDate'] =
```

```
df_collisions['OccurrenceDate'].dt.tz_localize(None).dt.round('H')
C:\Users\manav\AppData\Local\Temp\ipykernel_5196\2960273254.py:245:
FutureWarning: 'H' is deprecated and will be removed in a future version, please
use 'h' instead.
  df_env['LOCAL_DATE'] = df_env['LOCAL_DATE'].dt.tz_localize(None).dt.round('H')
C:\Users\manav\AppData\Local\Temp\ipykernel_5196\2960273254.py:250:
FutureWarning: 'H' is deprecated and will be removed in a future version, please
use 'h' instead.
  df_era5['timestamp'] = df_era5['timestamp'].dt.tz_localize(None).dt.round('H')
```

[9]:

```python
################################################################################
# CONTINUATION: Basic EDA/Exploration in Plain Text
#
# Use this immediately after your preprocessing code. Simply call:
#   basic_exploration_console(df_ksi, df_tmc, df_collisions, df_env, df_era5)
# to see shapes, df.info(), null counts, and a 25-row random sample.
################################################################################

import io

def explore_dataframe_console(df, df_name, unique_subset=None, sample_size=25):
    """
    Prints plain-text EDA information to the console:
      - shape
      - df.info() (column types, non-null counts)
      - null counts
      - sample_size random unique rows
    """
    print(f"\n=== Exploring {df_name} ===")
    print("Shape:", df.shape)

    # Capture df.info() output
    print("\n--- df.info() ---")
    info_buffer = io.StringIO()
    df.info(buf=info_buffer)
    info_str = info_buffer.getvalue()
    print(info_str)

    # Null counts
    print("\n--- Null Counts ---")
    null_series = df.isna().sum()
    print(null_series.to_string())

    # Display random unique rows
    _df = df.drop_duplicates(subset=unique_subset) if unique_subset else df.
↪drop_duplicates()
    nrows = min(sample_size, len(_df))
```

```
    if nrows > 0:
        sample = _df.sample(nrows, random_state=42)
        print(f"\n--- {df_name}: {nrows} Random Unique Rows ---")
        print(sample.to_string(index=False))
    else:
        print(f"No rows found in {df_name} after dropping duplicates.")


def basic_exploration_console(df_ksi, df_tmc, df_collisions, df_env, df_era5):
    """
    Runs explore_dataframe_console() on each final DataFrame,
    printing the results to the console in a copyable plain-text format.
    """
    explore_dataframe_console(df_ksi, "KSI Dataset",␣
 ↪unique_subset=["lon","lat"], sample_size=25)
    explore_dataframe_console(df_tmc, "TMC Dataset",␣
 ↪unique_subset=["lon","lat"], sample_size=25)
    explore_dataframe_console(df_collisions, "Collisions Dataset",␣
 ↪unique_subset=["lon","lat"], sample_size=25)
    explore_dataframe_console(df_env, "Env Canada Dataset",␣
 ↪unique_subset=["lon","lat"], sample_size=25)
    explore_dataframe_console(df_era5, "ERA5 Dataset",␣
 ↪unique_subset=["lon","lat"], sample_size=25)


# Usage Example:
# After your main script finishes and you have df_ksi, df_tmc, df_collisions,␣
 ↪df_env, df_era5:
basic_exploration_console(df_ksi, df_tmc, df_collisions, df_env, df_era5)
```

```
=== Exploring KSI Dataset ===
Shape: (5571, 21)

--- df.info() ---
<class 'pandas.core.frame.DataFrame'>
Index: 5571 entries, 11303 to 16873
Data columns (total 21 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   DATE               5571 non-null   datetime64[ns]
 1   TIME               5571 non-null   int64
 2   ACCLASS            5571 non-null   object
 3   INJURY             3141 non-null   object
 4   FATAL_NO           376 non-null    float64
 5   IMPACTYPE          5567 non-null   object
 6   ROAD_CLASS         5515 non-null   object
```

```
 7   ACCLOC             5533 non-null   object
 8   TRAFFCTL           5568 non-null   object
 9   VISIBILITY         5553 non-null   object
 10  LIGHT              5571 non-null   object
 11  RDSFCOND           5548 non-null   object
 12  VEHTYPE            3406 non-null   object
 13  INVTYPE            5564 non-null   object
 14  AUTOMOBILE         5086 non-null   object
 15  PEDESTRIAN         2358 non-null   object
 16  CYCLIST            607 non-null    object
 17  NEIGHBOURHOOD_158  5571 non-null   object
 18  lon                5571 non-null   float64
 19  lat                5571 non-null   float64
 20  datetime           5571 non-null   datetime64[ns]
dtypes: datetime64[ns](2), float64(3), int64(1), object(15)
memory usage: 957.5+ KB


--- Null Counts ---
DATE                    0
TIME                    0
ACCLASS                 0
INJURY               2430
FATAL_NO             5195
IMPACTYPE               4
ROAD_CLASS             56
ACCLOC                 38
TRAFFCTL                3
VISIBILITY             18
LIGHT                   0
RDSFCOND               23
VEHTYPE              2165
INVTYPE                 7
AUTOMOBILE            485
PEDESTRIAN           3213
CYCLIST              4964
NEIGHBOURHOOD_158       0
lon                     0
lat                     0
datetime                0

--- KSI Dataset: 25 Random Unique Rows ---
      DATE  TIME          ACCLASS  INJURY  FATAL_NO             IMPACTYPE
ROAD_CLASS                ACCLOC             TRAFFCTL VISIBILITY
LIGHT RDSFCOND                      VEHTYPE           INVTYPE AUTOMOBILE
PEDESTRIAN CYCLIST            NEIGHBOURHOOD_158        lon       lat
datetime
2019-06-09  1812 Non-Fatal Injury   Major       NaN             SMV Other Major
```

```
Arterial       Non Intersection           No Control      Clear         Daylight
Dry   Automobile, Station Wagon            Driver        Yes        NaN     NaN
High Park-Swansea -79.474414 43.635577 2019-06-09 18:00:00
2016-07-07   654 Non-Fatal Injury   Major       NaN              Rear End  Major
Arterial       At Intersection       Traffic Signal      Clear         Daylight
Dry                  Motorcycle Motorcycle Driver        Yes        NaN     NaN
West Humber-Clairville -79.605426 43.739031 2016-07-07 07:00:00
2016-02-23  2146 Non-Fatal Injury   Minor       NaN Pedestrian Collisions  Minor
Arterial      Non Intersection           No Control      Clear Dark, artificial
Dry   Automobile, Station Wagon            Driver        Yes        Yes     NaN
Trinity-Bellwoods -79.422458 43.653796 2016-02-23 22:00:00
2018-10-26  1157 Non-Fatal Injury     NaN       NaN             Sideswipe  Major
Arterial       At Intersection       Traffic Signal      Clear         Daylight
Dry   Automobile, Station Wagon            Driver        Yes        NaN     NaN
North Riverdale -79.358829 43.676131 2018-10-26 12:00:00
2016-04-10  1130 Non-Fatal Injury     NaN       NaN Pedestrian Collisions  Major
Arterial      Non Intersection Streetcar (Stop for)      Clear         Daylight
Dry   Automobile, Station Wagon            Driver        Yes        Yes     NaN
Yonge-Bay Corridor -79.386615 43.650843 2016-04-10 12:00:00
2016-08-21   634 Non-Fatal Injury   Major       NaN             SMV Other  Major
Arterial  Intersection Related           No Control      Clear         Daylight
Dry   Automobile, Station Wagon            Driver        Yes        NaN     NaN
Don Valley Village -79.363033 43.790763 2016-08-21 07:00:00
2018-01-02  1900 Non-Fatal Injury     NaN       NaN           Approaching  Major
Arterial      Non Intersection           No Control      Clear Dark, artificial
Dry               Passenger Van            Driver        Yes        NaN     NaN
Don Valley Village -79.362178 43.790946 2018-01-02 19:00:00
2016-04-06  1440 Non-Fatal Injury     NaN       NaN Pedestrian Collisions  Minor
Arterial       At Intersection           No Control      Clear         Daylight
Dry Municipal Transit Bus (TTC)            Driver        NaN        Yes     NaN
Stonegate-Queensway -79.503116 43.627577 2016-04-06 15:00:00
2016-09-20  1721 Non-Fatal Injury     NaN       NaN Pedestrian Collisions  Major
Arterial       At Intersection Pedestrian Crossover      Clear         Daylight
Dry   Automobile, Station Wagon            Driver        Yes        Yes     NaN
Runnymede-Bloor West Village -79.485935 43.652877 2016-09-20 17:00:00
2016-07-10  2240 Non-Fatal Injury     NaN       NaN Pedestrian Collisions  Major
Arterial       At Intersection       Traffic Signal      Clear Dark, artificial
Dry   Automobile, Station Wagon            Driver        Yes        Yes     NaN
Humber Bay Shores -79.481197 43.622859 2016-07-10 23:00:00
2017-04-23  2044 Non-Fatal Injury     NaN       NaN      Turning Movement  Major
Arterial       At Intersection   Traffic Controller      Clear            Dusk
Dry               Pick Up Truck            Driver        Yes        NaN     NaN
North Toronto -79.399272 43.711005 2017-04-23 21:00:00
2018-12-04  1932            Fatal     NaN       NaN Pedestrian Collisions  Major
Arterial      Non Intersection           No Control      Clear            Dark
Dry   Automobile, Station Wagon            Driver        Yes        Yes     NaN
Bendale-Glen Andrew -79.267612 43.758383 2018-12-04 20:00:00
2016-10-27  1506 Non-Fatal Injury     NaN       NaN Pedestrian Collisions  Major
```

```
Arterial      Non Intersection           No Control       Rain         Daylight
Wet   Automobile, Station Wagon            Driver        Yes        Yes     NaN
York University Heights -79.493770 43.781680 2016-10-27 15:00:00
2016-08-05  1225            Fatal   Fatal      50.0 Pedestrian Collisions  Major
Arterial      Non Intersection           No Control      Clear         Daylight
Dry                         NaN        Pedestrian        Yes        Yes     NaN
Willowdale West -79.422359 43.778243 2016-08-05 12:00:00
2017-10-18   906            Fatal   Fatal      50.0    Cyclist Collisions  Major
Arterial       At Intersection            Stop Sign      Clear         Daylight
Dry                     Bicycle           Cyclist        Yes        NaN     Yes
South Parkdale -79.431827 43.637950 2017-10-18 09:00:00
2016-12-21   959 Non-Fatal Injury     NaN       NaN Pedestrian Collisions  Major
Arterial       At Intersection       Traffic Signal      Clear         Daylight
Wet               Truck-Tractor      Truck Driver        NaN        Yes     NaN
Downsview -79.481548 43.724372 2016-12-21 10:00:00
2018-02-17  2319            Fatal   Fatal      10.0             SMV Other  Major
Arterial      Non Intersection           No Control        NaN             Dark
NaN   Automobile, Station Wagon            Driver        Yes        NaN     NaN
Bayview Woods-Steeles -79.392772 43.788287 2018-02-17 23:00:00
2015-06-07  1830 Non-Fatal Injury   Major       NaN    Cyclist Collisions  Major
Arterial       At Intersection       Traffic Signal      Clear         Daylight
Dry                     Bicycle           Cyclist        NaN        NaN     Yes
Lawrence Park South -79.402156 43.725032 2015-06-07 18:00:00
2019-12-13  2041 Non-Fatal Injury     NaN       NaN Pedestrian Collisions  Major
Arterial       At Intersection       Traffic Signal      Clear Dark, artificial
Dry   Automobile, Station Wagon            Driver        Yes        Yes     NaN
Downsview -79.486408 43.744704 2019-12-13 21:00:00
2016-05-19  1200 Non-Fatal Injury   Minor       NaN      Turning Movement  Major
Arterial       At Intersection       Traffic Signal      Clear         Daylight
Dry               Pick Up Truck            Driver        Yes        NaN     NaN
Rustic -79.505931 43.710971 2016-05-19 12:00:00
2017-06-13  2001 Non-Fatal Injury   Major       NaN             Sideswipe
NaN      Non Intersection           No Control      Clear         Daylight
Dry                  Motorcycle Motorcycle Driver        Yes        NaN     NaN
Banbury-Don Mills -79.330257 43.725211 2017-06-13 20:00:00
2017-10-23   833 Non-Fatal Injury   Major       NaN Pedestrian Collisions  Major
Arterial       At Intersection       Traffic Signal      Clear         Daylight
Dry                         NaN        Pedestrian        Yes        Yes     NaN
Wexford/Maryvale -79.310153 43.758959 2017-10-23 09:00:00
2018-10-04  1630 Non-Fatal Injury     NaN       NaN Pedestrian Collisions  Major
Arterial      Non Intersection           No Control      Clear         Daylight
Dry                Truck - Open      Truck Driver        NaN        Yes     NaN
Humber Summit -79.560573 43.767333 2018-10-04 16:00:00
2017-04-05  1454 Non-Fatal Injury     NaN       NaN Pedestrian Collisions
Local At/Near Private Drive           No Control      Clear         Daylight
Dry   Automobile, Station Wagon            Driver        Yes        Yes     NaN
Humber Heights-Westmount -79.522189 43.688971 2017-04-05 15:00:00
2019-09-17  2000            Fatal Minimal       NaN Pedestrian Collisions  Major
```

```
Arterial  Intersection Related            Stop Sign      Clear         Daylight
Dry               Pick Up Truck            Driver        Yes        Yes     NaN
Wexford/Maryvale -79.291835 43.745886 2019-09-17 20:00:00

=== Exploring TMC Dataset ===
Shape: (122570, 6)

--- df.info() ---
<class 'pandas.core.frame.DataFrame'>
Index: 122570 entries, 99531 to 223816
Data columns (total 6 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   count_date     122570 non-null  datetime64[ns]
 1   start_time     122570 non-null  datetime64[ns]
 2   end_time       122570 non-null  object
 3   lon            122570 non-null  float64
 4   lat            122570 non-null  float64
 5   location_name  122570 non-null  object
dtypes: datetime64[ns](2), float64(2), object(2)
memory usage: 6.5+ MB


--- Null Counts ---
count_date       0
start_time       0
end_time         0
lon              0
lat              0
location_name    0

--- TMC Dataset: 25 Random Unique Rows ---
count_date          start_time            end_time        lon       lat
location_name
2015-09-16 2015-09-16 08:00:00 2015-09-16T07:45:00 -79.375874 43.798753
Pineway Blvd / Cummer Ave
2017-07-05 2017-07-05 08:00:00 2017-07-05T07:45:00 -79.392200 43.713786
Mount Pleasant Rd / Keewatin Ave
2017-01-17 2017-01-17 08:00:00 2017-01-17T07:45:00 -79.383585 43.688320
Mount Pleasant Rd / Inglewood Dr
2018-11-14 2018-11-14 08:00:00 2018-11-14T07:45:00 -79.394934 43.690181
Yonge St / Heath St W
2018-03-21 2018-03-21 08:00:00 2018-03-21T07:45:00 -79.449648 43.641865
Parkdale Rd / Sunnyside Ave
2015-07-09 2015-07-09 08:00:00 2015-07-09T07:45:00 -79.363002 43.660087
Dundas St E / Sackville St
2016-07-20 2016-07-20 08:00:00 2016-07-20T07:45:00 -79.382783 43.670153
Church St / Hayden St
```

```
2017-07-19 2017-07-19 08:00:00 2017-07-19T07:45:00 -79.421189 43.699427
Glenayr Rd / Dewbourne Ave
2015-11-12 2015-11-12 08:00:00 2015-11-12T07:45:00 -79.524995 43.596419
Lake Shore Blvd W / Twenty Seventh St
2017-12-12 2017-12-12 08:00:00 2017-12-12T07:45:00 -79.512485 43.647118
Bloor St W / Cliveden Ave
2017-09-07 2017-09-07 08:00:00 2017-09-07T07:45:00 -79.453397 43.707908
Dufferin St / Glen Park Ave
2018-06-28 2018-06-28 08:00:00 2018-06-28T07:45:00 -79.409229 43.677112
Walmer Rd / Davenport Rd
2019-06-13 2019-06-13 08:00:00 2019-06-13T07:45:00 -79.404968 43.789742
Cummer Ave: Becky Cheung Crt - Willow Heights Crt
2015-06-15 2015-06-15 08:00:00 2015-06-15T07:45:00 -79.284073 43.693941
Denton Ave / Pharmacy Ave
2016-04-04 2016-04-04 08:00:00 2016-04-04T07:45:00 -79.417824 43.788978
Yonge St / Wedgewood Dr
2019-02-20 2019-02-20 08:00:00 2019-02-20T07:45:00 -79.309707 43.730014
Sloane Ave / Elvaston Dr
2018-02-08 2018-02-08 08:00:00 2018-02-08T07:45:00 -79.393036 43.716031
Mount Pleasant Rd / Sheldrake Blvd
2016-04-28 2016-04-28 08:00:00 2016-04-28T07:45:00 -79.428614 43.642218
Queen St W / Dufferin St
2016-12-15 2016-12-15 08:00:00 2016-12-15T07:45:00 -79.257147 43.758958
Brimley Rd / Dorcot Ave
2015-09-15 2015-09-15 08:00:00 2015-09-15T07:45:00 -79.389531 43.727791 Lawrence
Ave E / Wanless Cres / Ww E Wanless N Lawrence
2016-06-27 2016-06-27 08:00:00 2016-06-27T07:45:00 -79.488137 43.647728
Bloor St W / Riverside Dr
2018-04-03 2018-04-03 08:00:00 2018-04-03T07:45:00 -79.505567 43.709245
Jane St / Maple Leaf Dr / Church St
2016-01-04 2016-01-04 08:00:00 2016-01-04T07:45:00 -79.327686 43.760666
York Mills Rd / Fenside Dr
2016-11-08 2016-11-08 08:00:00 2016-11-08T07:45:00 -79.384759 43.672176
Church St / Park Rd
2016-05-03 2016-05-03 08:00:00 2016-05-03T07:45:00 -79.577138 43.705247
Martin Grove Rd / Vulcan St

=== Exploring Collisions Dataset ===
Shape: (417795, 8)

--- df.info() ---
<class 'pandas.core.frame.DataFrame'>
Index: 417795 entries, 9000 to 499537
Data columns (total 8 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   OccurrenceDate     417795 non-null  datetime64[ns]
 1   Hour               417795 non-null  int64
```

```
 2   lon                417795 non-null  float64
 3   lat                417795 non-null  float64
 4   Neighbourhood      417795 non-null  object
 5   Fatalities         417795 non-null  int64
 6   Injury_Collisions  417795 non-null  object
 7   PD_Collisions      417795 non-null  object
dtypes: datetime64[ns](1), float64(2), int64(2), object(3)
memory usage: 28.7+ MB


--- Null Counts ---
OccurrenceDate       0
Hour                 0
lon                  0
lat                  0
Neighbourhood        0
Fatalities           0
Injury_Collisions    0
PD_Collisions        0

--- Collisions Dataset: 25 Random Unique Rows ---
     OccurrenceDate  Hour        lon       lat
Neighbourhood  Fatalities Injury_Collisions PD_Collisions
2017-10-15 04:00:00    10 -79.487302 43.733588                Downsview-Roding-
CFB (26)           0                NO           YES
2020-02-13 05:00:00    13 -79.320056 43.799489
LAmoreaux (117)           0                NO            NO
2015-08-31 04:00:00    17 -79.634896 43.747439               West Humber-
Clairville (1)           0               YES            NO
2018-06-02 04:00:00    15 -79.194539 43.764873                           West
Hill (136)           0                NO           YES
2015-03-09 04:00:00    19 -79.356810 43.807030                   Hillcrest
Village (48)           0               YES            NO
2019-11-24 05:00:00    22 -79.580363 43.741743         Thistletown-Beaumond
Heights (3)           0                NO            NO
2018-07-04 04:00:00    16 -79.396341 43.723959                Lawrence Park
South (103)           0                NO           YES
2017-01-02 05:00:00    17 -79.409185 43.644676
Niagara (82)           0                NO            NO
2016-09-24 04:00:00    12 -79.413760 43.653422                    Trinity-
Bellwoods (81)           0                NO            NO
2017-07-22 04:00:00    14 -79.456588 43.683108                 Caledonia-
Fairbank (109)           0                NO           YES
2015-11-05 05:00:00     8 -79.548476 43.669889                   Princess-
Rosethorn (10)           0               YES            NO
2017-06-22 04:00:00    14 -79.493278 43.610738 Mimico (includes Humber Bay
Shores) (17)           0                NO           YES
2016-11-07 05:00:00    19 -79.452385 43.665497 Dovercourt-Wallace Emerson-
```

```
Junction (93)           0               YES            NO
2020-02-08 05:00:00    18 -79.337039 43.798128                        Pleasant
View (46)           0                NO            NO
2016-12-15 05:00:00    14 -79.326746 43.813126
Steeles (116)           0                NO           YES
2015-12-05 05:00:00    10 -79.394170 43.745935   Bridle Path-Sunnybrook-York
Mills (41)           0                NO           YES
2015-08-29 04:00:00    15 -79.407209 43.711169                      Yonge-
Eglinton (100)           0                NO           YES
2016-05-21 04:00:00    14 -79.379024 43.706246                 Mount Pleasant
East (99)           0                NO           YES
2017-05-23 04:00:00    14 -79.406685 43.744150                St.Andrew-
Windfields (40)           0                NO            NO
2020-09-30 04:00:00    15 -79.417985 43.669273
Annex (95)           0                NO           YES
2015-05-01 04:00:00    10 -79.436326 43.725919                 Englemount-
Lawrence (32)           0                NO           YES
2017-01-24 05:00:00    16 -79.445699 43.664897 Dovercourt-Wallace Emerson-
Junction (93)           0                NO           YES
2015-10-29 04:00:00    16 -79.327524 43.726645                      Flemingdon
Park (44)           0                NO           YES
2015-11-02 05:00:00     7 -79.471476 43.690978            Beechborough-
Greenbrook (112)           0                NO           YES
2015-01-10 05:00:00     4 -79.328359 43.729355                   Banbury-Don
Mills (42)           0                NO            NO

=== Exploring Env Canada Dataset ===
Shape: (52355, 10)

--- df.info() ---
<class 'pandas.core.frame.DataFrame'>
Index: 52355 entries, 36750 to 89104
Data columns (total 10 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   LOCAL_DATE         52355 non-null  datetime64[ns]
 1   LOCAL_HOUR         52355 non-null  int64
 2   TEMP               52345 non-null  float64
 3   WINDCHILL          0 non-null      float64
 4   PRECIP_AMOUNT      51802 non-null  float64
 5   RELATIVE_HUMIDITY  52347 non-null  float64
 6   VISIBILITY         0 non-null      float64
 7   WEATHER_ENG_DESC   0 non-null      float64
 8   lon                52355 non-null  float64
 9   lat                52355 non-null  float64
dtypes: datetime64[ns](1), float64(8), int64(1)
memory usage: 4.4 MB
```

```
--- Null Counts ---
LOCAL_DATE               0
LOCAL_HOUR               0
TEMP                    10
WINDCHILL            52355
PRECIP_AMOUNT          553
RELATIVE_HUMIDITY        8
VISIBILITY           52355
WEATHER_ENG_DESC     52355
lon                      0
lat                      0

--- Env Canada Dataset: 1 Random Unique Rows ---
LOCAL_DATE  LOCAL_HOUR  TEMP  WINDCHILL  PRECIP_AMOUNT  RELATIVE_HUMIDITY
VISIBILITY  WEATHER_ENG_DESC   lon       lat
2020-12-31           0   3.2        NaN            0.0               63.0
NaN               NaN -79.4 43.666667

=== Exploring ERA5 Dataset ===
Shape: (525850, 8)

--- df.info() ---
<class 'pandas.core.frame.DataFrame'>
Index: 525850 entries, 0 to 526056
Data columns (total 8 columns):
 #   Column                   Non-Null Count   Dtype
---  ------                   --------------   -----
 0   timestamp                525850 non-null  datetime64[ns]
 1   temperature_2m           525850 non-null  float64
 2   dewpoint_temperature_2m  525850 non-null  float64
 3   total_precipitation      525850 non-null  float64
 4   u_component_of_wind_10m  525850 non-null  float64
 5   v_component_of_wind_10m  525850 non-null  float64
 6   lon                      525850 non-null  float64
 7   lat                      525850 non-null  float64
dtypes: datetime64[ns](1), float64(7)
memory usage: 36.1 MB


--- Null Counts ---
timestamp                  0
temperature_2m             0
dewpoint_temperature_2m    0
total_precipitation        0
u_component_of_wind_10m    0
v_component_of_wind_10m    0
lon                        0
```

```
lat                         0

--- ERA5 Dataset: 10 Random Unique Rows ---
 timestamp  temperature_2m  dewpoint_temperature_2m  total_precipitation
u_component_of_wind_10m  v_component_of_wind_10m      lon     lat
2015-01-01      267.469437               256.379730            0.000267
7.538605                 3.411713 -79.3641 43.7326
2015-01-01      267.221390               255.830902            0.000185
6.333527                 3.142181 -79.5181 43.7731
2015-01-01      268.500687               259.366058            0.000525
10.322784                 4.029877 -79.1845 43.7636
2015-01-01      267.469437               256.379730            0.000267
7.538605                 3.411713 -79.3832 43.6532
2015-01-01      268.500687               259.366058            0.000525
10.322784                 4.029877 -79.2454 43.7078
2015-01-01      268.500687               259.366058            0.000525
10.322784                 4.029877 -79.2263 43.7845
2015-01-01      267.162796               256.330902            0.000333
7.084503                 3.222260 -79.2939 43.7996
2015-01-01      267.738968               257.090668            0.000398
8.947784                 3.726166 -79.3017 43.6650
2015-01-01      266.846390               255.664886            0.000197
5.579620                 2.945892 -79.6205 43.5906
2015-01-01      267.469437               256.379730            0.000267
7.538605                 3.411713 -79.4309 43.6816
```

[ ]: