Wrangling

December 1, 2023

1 Data Wrangling of Typhoon Incidences

1.1 Assessment

[ ]:
```
# Import required libraries
import pandas as pd
import numpy as np

# Load the original DataFrame from a CSV file
df_original = pd.read_csv('data_typhoons.txt')

df_original.head()
```

[ ]:
```
   Unnamed: 0.2  Unnamed: 0.1  Unnamed: 0  \
0             0             0           0
1             1             1           1
2             2             2           2
3             3             3           3
4             4             4           4

  66666 5101   10      5101 0 6                                    19901017
0  51021906 002 2 200 1385 1010                 …
1  51021912 002 2 200 1385 1010                 …
2  51021918 002 2 230 1421 1000                 …
3  51022000 002 9 250 1460  994                 …
4  51022006 002 9 276 1506  994                 …
```

[ ]:
```
# Define columns to drop from the original DataFrame
columns_to_drop = ['Unnamed: 0.2', 'Unnamed: 0.1', 'Unnamed: 0']

# Drop the specified columns from the original DataFrame
df_columns_dropped = df_original.drop(columns_to_drop, axis=1)

df_columns_dropped.head()
```

[ ]:
```
  66666 5101   10      5101 0 6                                    19901017
0  51021906 002 2 200 1385 1010                 …
1  51021912 002 2 200 1385 1010                 …
```

```
2  51021918 002 2 230 1421 1000                    …
3  51022000 002 9 250 1460  994                    …
4  51022006 002 9 276 1506  994                    …
```

1.2 Splitting

[ ]:
```
# Store the column to be split
column_to_split = df_columns_dropped.columns

# Split the specified columns based on spaces and concatenate them with the␣
 ↪original DataFrame
for column in column_to_split:

    # Split on whitespaces creating new columns
    split_df = df_columns_dropped[column].str.split(expand=True)

    # Create names for the new columns
    split_df.columns = [f'{column}_{i + 1}' for i in range(split_df.shape[1])]

    # Concatenate the split DataFrame with the original DataFrame
    df_columns_dropped = pd.concat([df_columns_dropped, split_df], axis=1)

    # Drop the original column
    df_columns_dropped.drop(column, axis=1, inplace=True)

df_columns_dropped.head()
```

[ ]:
```
  66666 5101   10      5101 0 6                                19901017
_1  \
0                                        51021906
1                                        51021912
2                                        51021918
3                                        51022000
4                                        51022006

  66666 5101   10      5101 0 6                                19901017
_2  \
0                                            002
1                                            002
2                                            002
3                                            002
4                                            002

  66666 5101   10      5101 0 6                                19901017
_3  \
0                                              2
1                                              2
```

```
2                                                    2
3                                                    9
4                                                    9

  66666 5101   10      5101 0 6                                  19901017
_4  \
0                                                  200
1                                                  200
2                                                  230
3                                                  250
4                                                  276

  66666 5101   10      5101 0 6                                  19901017
_5  \
0                                                 1385
1                                                 1385
2                                                 1421
3                                                 1460
4                                                 1506

  66666 5101   10      5101 0 6                                  19901017
_6  \
0                                                 1010
1                                                 1010
2                                                 1000
3                                                  994
4                                                  994

  66666 5101   10      5101 0 6                                  19901017
_7  \
0                                                 None
1                                                 None
2                                                 None
3                                                 None
4                                                 None

  66666 5101   10      5101 0 6                                  19901017
_8  \
0                                                 None
1                                                 None
2                                                 None
3                                                 None
4                                                 None

  66666 5101   10      5101 0 6                                  19901017
_9  \
0                                                 None
```

```
1                                                    None
2                                                    None
3                                                    None
4                                                    None

  66666 5101   10      5101 0 6                                  19901017
_10  \
0                                                    None
1                                                    None
2                                                    None
3                                                    None
4                                                    None

  66666 5101   10      5101 0 6                                  19901017
_11  \
0                                                    None
1                                                    None
2                                                    None
3                                                    None
4                                                    None

  66666 5101   10      5101 0 6                                  19901017
_12
0                                                    None
1                                                    None
2                                                    None
3                                                    None
4                                                    None
```

1.3 Renaming

[ ]:
```
# Based on the data's dictionary define the list of columns to be created based␣
 ↪on splitting
columns_to_be_created = [
    'Date and Time',
    'Indicator',
    'Grade',
    'Latitude of the Center',
    'Longitude of the Center',
    'Central Pressure',
    'Maximum sustained wind speed',
    'Direction of the longest radius of 50kt winds or greater',
    'Longest radius of 50kt winds or greater',
    'Shortest radius of 50kt winds or greater',
    'Direction of the longest radius of 30kt winds or greater',
    'Longest radius of 30kt winds or greater',
    'Shortest radius of 30kt winds or greater',
```

```
    'Indicator of landfall or passage'
]

# Rename the columns to match the desired names
df_columns_dropped.rename(columns=dict(zip(df_columns_dropped.columns,␣
 ↪columns_to_be_created)), inplace=True)

df_columns_dropped.head()
```

[ ]:

	Date and Time	Indicator	Grade	Latitude of the Center	Longitude of the Center	Central Pressure	Maximum sustained wind speed	Direction of the longest radius of 50kt winds or greater	Longest radius of 50kt winds or greater	Shortest radius of 50kt winds or greater	Direction of the longest radius of 30kt winds or greater
0	51021906	002	2	200	1385	1010	None	None	None	None	None
1	51021912	002	2	200	1385	1010	None	None	None	None	None
2	51021918	002	2	230	1421	1000	None	None	None	None	None
3	51022000	002	9	250	1460	994	None	None	None	None	None
4	51022006	002	9	276	1506	994	None	None	None	None	None

```
2                                                    None
3                                                    None
4                                                    None

  Longest radius of 30kt winds or greater
0                                       None
1                                       None
2                                       None
3                                       None
4                                       None
```

1.4 Formatting

[ ]:
```
# Convert columns to numeric ignoring errors
columns_to_numeric = df_columns_dropped.columns
for column in columns_to_numeric:
    df_columns_dropped[column] = pd.to_numeric(df_columns_dropped[column],␣
 ↪errors='ignore')
```

1.5 Identifying

[ ]:
```
# Fill missing values in the 'Maximum sustained wind speed' column with '0'
df_columns_dropped['Maximum sustained wind speed'].fillna('0', inplace=True)

# Initialize lists to store storm's names, IDs and positions
names = []
international_id = []
position = []

# Set counters
rows = len(df_columns_dropped) - 1
row = 0

# Iterate through the dataframe
while row < rows:

    # Check for different storms
    if df_columns_dropped.iloc[row, 0] == 66666:

        # Append storm's names, IDs and positions
        names.append(df_columns_dropped.loc[row, 'Maximum sustained wind␣
 ↪speed'])
        international_id.append(df_columns_dropped.loc[row, 'Indicator'])
        position.append(row)
    row = row + 1
```

[ ]:
```
# Set indexes that will guide the data to be stored
lower_index = 0
upper_index = 1

# Create lists for names and IDs ot be set as columns' names
column_names = []
international_id_numbers = []

# Run the index through the dataframe
while upper_index < len(position):

    # Set the amount of entries are presented for each storm
    column_names = column_names + ((position[upper_index] -␣
 ↪position[lower_index]) - 1) * [names[lower_index]]
    international_id_numbers = international_id_numbers +␣
 ↪((position[upper_index] - position[lower_index]) - 1) *␣
 ↪[international_id[lower_index]]
    lower_index = lower_index + 1
    upper_index = upper_index + 1
```

[ ]:
```
# Drop rows based on positions and extra rows
df_rows_dropped = df_columns_dropped.drop(position)
extra_rows = list(range(68697, 68742))
df_rows_dropped = df_rows_dropped.drop(extra_rows, errors='ignore')

# Create a list of storm names for the remaining rows
storm_names = 10 * ['0'] + column_names
id_numbers = 10 * ['5101'] + international_id_numbers

# Assign the 'storm_name' column to the DataFrame
df_organized = df_rows_dropped.
 ↪assign(storm_names=storm_names,id_numbers=id_numbers)

# Rename columns
df_organized.rename(columns={'storm_names':'Storm Names','id_numbers':
 ↪'International ID'}, inplace=True)

df_organized
```

[ ]:

	Date and Time	Indicator	Grade	Latitude of the Center	\
0	51021906	2	2	200	
1	51021912	2	2	200	
2	51021918	2	2	230	
3	51022000	2	9	250	
4	51022006	2	9	276	
…	…	…	…	…	
70628	23061518	2	6	393	

```
70629       23061600          2      6                    392
70630       23061606          2      6                    396
70631       23061612          2      6                    396
70632       23061618          2      6                    397

       Longitude of the Center  Central Pressure Maximum sustained wind speed  \
0                         1385              1010                            0
1                         1385              1010                            0
2                         1421              1000                            0
3                         1460               994                            0
4                         1506               994                            0
…                          …                 …                              …
70628                     1745               990                          000
70629                     1765               986                          000
70630                     1781               984                          000
70631                     1796               984                          000
70632                     1809               986                          000

      Direction of the longest radius of 50kt winds or greater  \
0                                                        None
1                                                        None
2                                                        None
3                                                        None
4                                                        None
…                                                         …
70628                                                    None
70629                                                    None
70630                                                    None
70631                                                    None
70632                                                    None

       Longest radius of 50kt winds or greater  \
0                                          NaN
1                                          NaN
2                                          NaN
3                                          NaN
4                                          NaN
…                                           …
70628                                      NaN
70629                                      NaN
70630                                      NaN
70631                                      NaN
70632                                      NaN

       Shortest radius of 50kt winds or greater  \
0                                           NaN
1                                           NaN
```

```
2                                                  NaN
3                                                  NaN
4                                                  NaN
…                                                  …
70628                                              NaN
70629                                              NaN
70630                                              NaN
70631                                              NaN
70632                                              NaN

       Direction of the longest radius of 30kt winds or greater  \
0                                                        NaN
1                                                        NaN
2                                                        NaN
3                                                        NaN
4                                                        NaN
…                                                        …
70628                                                    NaN
70629                                                    NaN
70630                                                    NaN
70631                                                    NaN
70632                                                    NaN

      Longest radius of 30kt winds or greater Storm Names International ID
0                                        None           0             5101
1                                        None           0             5101
2                                        None           0             5101
3                                        None           0             5101
4                                        None           0             5101
…                                        …              …             …
70628                                    None           6             2302
70629                                    None           6             2302
70630                                    None           6             2302
70631                                    None           6             2302
70632                                    None           6             2302

[68706 rows x 14 columns]
```

[ ]:
```
df_organized.loc[60,'Storm Names']
```

[ ]:
```
'HOPE'
```

1.6 Reordering

[ ]:
```
# Define the order of columns in the final DataFrame
column_order = [
    'International ID',
    'Storm Names',
    'Date and Time',
    'Indicator',
    'Grade',
    'Latitude of the Center',
    'Longitude of the Center',
    'Central Pressure',
    'Maximum sustained wind speed',
    'Direction of the longest radius of 50kt winds or greater',
    'Longest radius of 50kt winds or greater',
    'Shortest radius of 50kt winds or greater',
    'Direction of the longest radius of 30kt winds or greater',
    'Longest radius of 30kt winds or greater'
]

# Reorder columns
df_ordered = df_organized[column_order]

# Replace '0' with NaN
df_ordered.replace({'0': np.nan}, inplace=True)

df_ordered.head()
```

[ ]:
```
   International ID Storm Names  Date and Time  Indicator  Grade  \
0              5101         NaN       51021906          2      2
1              5101         NaN       51021912          2      2
2              5101         NaN       51021918          2      2
3              5101         NaN       51022000          2      9
4              5101         NaN       51022006          2      9

   Latitude of the Center  Longitude of the Center  Central Pressure  \
0                     200                     1385              1010
1                     200                     1385              1010
2                     230                     1421              1000
3                     250                     1460               994
4                     276                     1506               994

  Maximum sustained wind speed  \
0                          NaN
1                          NaN
2                          NaN
3                          NaN
```

```
4                          NaN

  Direction of the longest radius of 50kt winds or greater  \
0                                                      None
1                                                      None
2                                                      None
3                                                      None
4                                                      None

   Longest radius of 50kt winds or greater  \
0                                      NaN
1                                      NaN
2                                      NaN
3                                      NaN
4                                      NaN

   Shortest radius of 50kt winds or greater  \
0                                       NaN
1                                       NaN
2                                       NaN
3                                       NaN
4                                       NaN

   Direction of the longest radius of 30kt winds or greater  \
0                                                      NaN
1                                                      NaN
2                                                      NaN
3                                                      NaN
4                                                      NaN

  Longest radius of 30kt winds or greater
0                                    None
1                                    None
2                                    None
3                                    None
4                                    None
```

1.7 Classifying

[ ]:
```
# Map numerical grade values to their corresponding descriptions
map_grade = {
    2: 'Tropical Depression (TD)',
    3: 'Tropical Storm (TS)',
    4: 'Severe Tropical Storm (STS)',
    5: 'Typhoon (TY)',
    6: 'Extra-tropical Cyclone (L)',
    7: 'Just entering into the responsible area of RSMC Tokyo-Typhoon Center',
```

```
    8: 'Not used',
    9: 'Tropical Cyclone of TS intensity or higher',
}

# Replace numerical grade values with their descriptions
df_ordered['Grade'] = df_ordered['Grade'].replace(map_grade)

# Drop the 'Indicator' column
df_categorized = df_ordered.drop(columns='Indicator')

df_categorized.head()
```

[ ]:

	International ID	Storm Names	Date and Time	Grade	Latitude of the Center	Longitude of the Center	Central Pressure	Maximum sustained wind speed	Direction of the longest radius of 50kt winds or greater	Longest radius of 50kt winds or greater
0	5101	NaN	51021906	Tropical Depression (TD)	200	1385	1010	NaN	None	NaN
1	5101	NaN	51021912	Tropical Depression (TD)	200	1385	1010	NaN	None	NaN
2	5101	NaN	51021918	Tropical Depression (TD)	230	1421	1000	NaN	None	NaN
3	5101	NaN	51022000	Tropical Cyclone of TS intensity or higher	250	1460	994	NaN	None	NaN
4	5101	NaN	51022006	Tropical Cyclone of TS intensity or higher	276	1506	994	NaN	None	NaN

```
   Shortest radius of 50kt winds or greater  \
0                                        NaN
1                                        NaN
2                                        NaN
3                                        NaN
4                                        NaN

   Direction of the longest radius of 30kt winds or greater  \
0                                                        NaN
1                                                        NaN
2                                                        NaN
3                                                        NaN
4                                                        NaN

  Longest radius of 30kt winds or greater
0                                    None
1                                    None
2                                    None
3                                    None
4                                    None
```

1.8 Parsing

[ ]:
```python
# Define a custom function to parse date strings
def parse_custom_datetime(input_string):
    try:
        # Extract year, month, day, and hour from the input string
        year = input_string[:2]
        month = input_string[2:4]
        day = input_string[4:6]
        hour = input_string[6:8]

        # Format the components into the desired format
        formatted_datetime = f"{year}-{month}-{day} {hour}:00:00"
        return formatted_datetime
    except IndexError:
        print("Error: Input string does not have the expected length.")
        return None

# Parse date strings using the custom function
i = 0
parsed = []
while i < len(df_categorized):
    input_string = str(df_categorized.iloc[i, 2])
    parsed_datetime = parse_custom_datetime(input_string)
    parsed.append(parsed_datetime)
    i = i + 1
```

```
# Add the parsed date as a new column
df_date_and_time = df_ordered.assign(Date_Time_Parsed=parsed)

# Drop the original 'Date and Time' column
df_date_and_time = df_date_and_time.drop(columns='Date and Time')

# Define a custom function to adjust two-digit years
def parse_two_digit_year(x):
    try:
        # Split the date components
        parts = x.split('-')

        # Convert the year part to an integer
        year = int(parts[0])

        # Adjust the year to be in the 20th century if necessary
        if year < 50:
            year += 2000
        else:
            year += 1900

        # Reconstruct the date with the adjusted year
        return f"{year}-{parts[1]}-{parts[2]}"
    except IndexError:
        print("Error: Input string does not have the expected length.")
        return None

# Apply the custom function to the 'Date' column
df_date_and_time['Parsed_Date_and_Time'] = df_date_and_time['Date_Time_Parsed'].
 ↪apply(parse_two_digit_year)

# Drop the 'Date_Time_Parsed' column
df_date_time_dropped = df_date_and_time.drop(columns='Date_Time_Parsed')

# Drop the original 'Date' column and rename the new 'Parsed_Date' column
df_date_time_dropped.rename(columns={'Parsed_Date_and_Time': 'Date and Time'},␣
 ↪inplace=True)

df_date_time_dropped.head()
```

[ ]:

	International ID	Storm Names	Indicator	\
0	5101	NaN	2	
1	5101	NaN	2	
2	5101	NaN	2	
3	5101	NaN	2	
4	5101	NaN	2	

```
                                    Grade  Latitude of the Center  \
0                   Tropical Depression (TD)                    200
1                   Tropical Depression (TD)                    200
2                   Tropical Depression (TD)                    230
3  Tropical Cyclone of TS intensity or higher                   250
4  Tropical Cyclone of TS intensity or higher                   276

   Longitude of the Center  Central Pressure Maximum sustained wind speed  \
0                     1385              1010                          NaN
1                     1385              1010                          NaN
2                     1421              1000                          NaN
3                     1460               994                          NaN
4                     1506               994                          NaN

  Direction of the longest radius of 50kt winds or greater  \
0                                                     None
1                                                     None
2                                                     None
3                                                     None
4                                                     None

   Longest radius of 50kt winds or greater  \
0                                      NaN
1                                      NaN
2                                      NaN
3                                      NaN
4                                      NaN

   Shortest radius of 50kt winds or greater  \
0                                       NaN
1                                       NaN
2                                       NaN
3                                       NaN
4                                       NaN

   Direction of the longest radius of 30kt winds or greater  \
0                                                       NaN
1                                                       NaN
2                                                       NaN
3                                                       NaN
4                                                       NaN

  Longest radius of 30kt winds or greater        Date and Time
0                                    None  1951-02-19 06:00:00
1                                    None  1951-02-19 12:00:00
2                                    None  1951-02-19 18:00:00
```

```
3                                    None  1951-02-20 00:00:00
4                                    None  1951-02-20 06:00:00
```

[ ]:
```
# Define the new order of columns
new_column_order = [
    'International ID',
    'Storm Names',
    'Date and Time',
    'Grade',
    'Latitude of the Center',
    'Longitude of the Center',
    'Central Pressure',
    'Maximum sustained wind speed',
    'Direction of the longest radius of 50kt winds or greater',
    'Longest radius of 50kt winds or greater',
    'Shortest radius of 50kt winds or greater',
    'Direction of the longest radius of 30kt winds or greater',
    'Longest radius of 30kt winds or greater'
]

# Reorder columns
df_new_order = df_date_time_dropped[new_column_order]

# Convert the 'Date' column to datetime format
df_new_order.loc[:,'Date and Time'] = pd.to_datetime(df_new_order['Date and␣
 ↪Time'], errors='coerce')

df_new_order
```

[ ]:
```
      International ID Storm Names        Date and Time  \
0                 5101         NaN  1951-02-19 06:00:00
1                 5101         NaN  1951-02-19 12:00:00
2                 5101         NaN  1951-02-19 18:00:00
3                 5101         NaN  1951-02-20 00:00:00
4                 5101         NaN  1951-02-20 06:00:00
…                  …           …                   …
70628             2302           6  2023-06-15 18:00:00
70629             2302           6  2023-06-16 00:00:00
70630             2302           6  2023-06-16 06:00:00
70631             2302           6  2023-06-16 12:00:00
70632             2302           6  2023-06-16 18:00:00

                                        Grade  Latitude of the Center  \
0                       Tropical Depression (TD)                     200
1                       Tropical Depression (TD)                     200
2                       Tropical Depression (TD)                     230
3     Tropical Cyclone of TS intensity or higher                     250
```

```
4      Tropical Cyclone of TS intensity or higher                      276
…                                                 …                  …
70628                 Extra-tropical Cyclone (L)                      393
70629                 Extra-tropical Cyclone (L)                      392
70630                 Extra-tropical Cyclone (L)                      396
70631                 Extra-tropical Cyclone (L)                      396
70632                 Extra-tropical Cyclone (L)                      397

       Longitude of the Center  Central Pressure Maximum sustained wind speed  \
0                         1385              1010                          NaN
1                         1385              1010                          NaN
2                         1421              1000                          NaN
3                         1460               994                          NaN
4                         1506               994                          NaN
…                          …                 …                             …
70628                     1745               990                          000
70629                     1765               986                          000
70630                     1781               984                          000
70631                     1796               984                          000
70632                     1809               986                          000

      Direction of the longest radius of 50kt winds or greater  \
0                                                    None
1                                                    None
2                                                    None
3                                                    None
4                                                    None
…                                                    …
70628                                                None
70629                                                None
70630                                                None
70631                                                None
70632                                                None

       Longest radius of 50kt winds or greater  \
0                                          NaN
1                                          NaN
2                                          NaN
3                                          NaN
4                                          NaN
…                                          …
70628                                      NaN
70629                                      NaN
70630                                      NaN
70631                                      NaN
70632                                      NaN
```

```
       Shortest radius of 50kt winds or greater  \
0                                           NaN
1                                           NaN
2                                           NaN
3                                           NaN
4                                           NaN
…                                           …
70628                                       NaN
70629                                       NaN
70630                                       NaN
70631                                       NaN
70632                                       NaN

       Direction of the longest radius of 30kt winds or greater  \
0                                                     NaN
1                                                     NaN
2                                                     NaN
3                                                     NaN
4                                                     NaN
…                                                     …
70628                                                 NaN
70629                                                 NaN
70630                                                 NaN
70631                                                 NaN
70632                                                 NaN

      Longest radius of 30kt winds or greater
0                                        None
1                                        None
2                                        None
3                                        None
4                                        None
…                                        …
70628                                    None
70629                                    None
70630                                    None
70631                                    None
70632                                    None

[68706 rows x 13 columns]
```

1.9 Exporting

[ ]:
```
df_new_order.to_csv('Cleaned.csv')
```