DSC540-Project Milestone4-Jyoti Dave

February 5, 2025

[1]:
```
# Weeks 9 & 10 Term project : Milestone 4
```

[2]:
```
# Connecting to an API/Pulling in the Data and Cleaning/Formatting
# Perform at least 5 data transformation and/or cleansing steps to your API␣
 ↪data.
# Examples:
# Replace Headers
# Format data into a more readable format
# Identify outliers and bad data
# Find duplicates
# Fix casing or inconsistent values
# Conduct Fuzzy Matching
```

[3]:
```
# API:
# •        Description: Covid tracking API, contains datasets such as state,␣
 ↪hospitalized, positive, negative case details by state wise in USA.
# •        Source: Covid tracking API.
# •        Link: https://api.covidtracking.com/v1/states/current.json
```

[4]:
```
import requests
import pandas as pd

# Fetch API data
url = "https://api.covidtracking.com/v1/states/current.json"
response = requests.get(url)
data = response.json()

# Convert JSON data to a DataFrame
df = pd.DataFrame(data)

# View raw data structure
print(df.head())

# Step 1. Rename columns for readability
df.rename(columns={
    "state": "State",
    "positive": "Positive_Cases",
```

```python
    "negative": "Negative_Cases",
    "pending": "Pending_Tests",
    "hospitalizedCurrently": "Currently_Hospitalized",
    "hospitalizedCumulative": "Total_Hospitalized",
    "inIcuCurrently": "Currently_in_ICU",
    "inIcuCumulative": "Total_ICU",
    "onVentilatorCurrently": "Currently_on_Ventilator",
    "onVentilatorCumulative": "Total_on_Ventilator",
    "recovered": "Recovered",
    "death": "Deaths",
    "totalTestResults": "Total_Tests",
    "lastUpdateEt": "Last_Update",
    "dateModified": "Date_Modified",
    "fips": "FIPS_Code"
}, inplace=True)

# View cleaned dataset
print(df.head())
```

```
       date state  positive  probableCases   negative  pending  \
0  20210307    AK     56886            NaN        NaN      NaN
1  20210307    AL    499819       107742.0  1931711.0      NaN
2  20210307    AR    324818        69092.0  2480716.0      NaN
3  20210307    AS         0            NaN     2140.0      NaN
4  20210307    AZ    826454        56519.0  3073010.0      NaN

  totalTestResultsSource  totalTestResults  hospitalizedCurrently  \
0        totalTestsViral           1731628                   33.0
1  totalTestsPeopleViral           2323788                  494.0
2        totalTestsViral           2736442                  335.0
3        totalTestsViral              2140                    NaN
4        totalTestsViral           7908105                  963.0

   hospitalizedCumulative  …  dataQualityGrade  deathIncrease  \
0                  1293.0  …              None              0
1                 45976.0  …              None             -1
2                 14926.0  …              None             22
3                     NaN  …              None              0
4                 57907.0  …              None              5

   hospitalizedIncrease                                      hash  \
0                     0  dc4bccd4bb885349d7e94d6fed058e285d4be164
1                     0  997207b430824ea40b8eb8506c19a93e07bc972e
2                    11  50921aeefba3e30d31623aa495b47fb2ecc72fae
3                     0  f77912d0b80d579fbb6202fa1a90554fc4dc1443
4                    44  0437a7a96f4471666f775e63e86923eb5cbd8cdf

   commercialScore negativeRegularScore negativeScore positiveScore  score  \
0                0                    0             0             0      0
1                0                    0             0             0      0
2                0                    0             0             0      0
3                0                    0             0             0      0
4                0                    0             0             0      0

   grade
0
1
2
3
4

[5 rows x 56 columns]
       date State  Positive_Cases  probableCases  Negative_Cases  \
0  20210307    AK           56886            NaN             NaN
1  20210307    AL          499819       107742.0       1931711.0
2  20210307    AR          324818        69092.0       2480716.0
3  20210307    AS               0            NaN          2140.0
4  20210307    AZ          826454        56519.0       3073010.0

   Pending_Tests totalTestResultsSource  Total_Tests  Currently_Hospitalized  \
0            NaN        totalTestsViral      1731628                    33.0
1            NaN  totalTestsPeopleViral      2323788                   494.0
2            NaN        totalTestsViral      2736442                   335.0
3            NaN        totalTestsViral         2140                     NaN
4            NaN        totalTestsViral      7908105                   963.0

   Total_Hospitalized  …  dataQualityGrade  deathIncrease  \
0              1293.0  …              None              0
1             45976.0  …              None             -1
2             14926.0  …              None             22
3                 NaN  …              None              0
4             57907.0  …              None              5

   hospitalizedIncrease                                      hash  \
0                     0  dc4bccd4bb885349d7e94d6fed058e285d4be164
1                     0  997207b430824ea40b8eb8506c19a93e07bc972e
2                    11  50921aeefba3e30d31623aa495b47fb2ecc72fae
3                     0  f77912d0b80d579fbb6202fa1a90554fc4dc1443
4                    44  0437a7a96f4471666f775e63e86923eb5cbd8cdf

   commercialScore negativeRegularScore negativeScore positiveScore  score  \
0                0                    0             0             0      0
1                0                    0             0             0      0
2                0                    0             0             0      0
3                0                    0             0             0      0
4                0                    0             0             0      0
```

```
   grade
0
1
2
3
4

[5 rows x 56 columns]
```

[5]:
```
# Renamed Columns - Made them more descriptive
```

[6]:
```
# Step 2. Standardize state abbreviations to uppercase
df["State"] = df["State"].str.upper()

# View cleaned dataset
print(df.head())
```

```
       date State  Positive_Cases  probableCases  Negative_Cases  \
0  20210307    AK           56886            NaN             NaN
1  20210307    AL          499819       107742.0       1931711.0
2  20210307    AR          324818        69092.0       2480716.0
3  20210307    AS               0            NaN          2140.0
4  20210307    AZ          826454        56519.0       3073010.0

   Pending_Tests totalTestResultsSource  Total_Tests  Currently_Hospitalized  \
0            NaN         totalTestsViral      1731628                    33.0
1            NaN  totalTestsPeopleViral      2323788                   494.0
2            NaN         totalTestsViral      2736442                   335.0
3            NaN         totalTestsViral         2140                     NaN
4            NaN         totalTestsViral      7908105                   963.0

   Total_Hospitalized  …  dataQualityGrade  deathIncrease  \
0              1293.0  …              None              0
1             45976.0  …              None             -1
2             14926.0  …              None             22
3                 NaN  …              None              0
4             57907.0  …              None              5

   hospitalizedIncrease                                      hash  \
0                     0  dc4bccd4bb885349d7e94d6fed058e285d4be164
1                     0  997207b430824ea40b8eb8506c19a93e07bc972e
2                    11  50921aeefba3e30d31623aa495b47fb2ecc72fae
3                     0  f77912d0b80d579fbb6202fa1a90554fc4dc1443
4                    44  0437a7a96f4471666f775e63e86923eb5cbd8cdf

   commercialScore negativeRegularScore negativeScore positiveScore  score  \
0                0                    0             0             0      0
```

```
1                0                    0             0             0      0
2                0                    0             0             0      0
3                0                    0             0             0      0
4                0                    0             0             0      0

   grade
0
1
2
3
4

[5 rows x 56 columns]
```

[7]:

```
# Standardized State Codes - Ensured all state abbreviations are in uppercase
```

[8]:

```
# Step 3. Convert date columns to datetime format
df["Last_Update"] = pd.to_datetime(df["Last_Update"], errors='coerce')
df["Date_Modified"] = pd.to_datetime(df["Date_Modified"], errors='coerce')

# View cleaned dataset
print(df.head())
```

```
       date State  Positive_Cases  probableCases  Negative_Cases  \
0  20210307    AK           56886            NaN             NaN
1  20210307    AL          499819       107742.0       1931711.0
2  20210307    AR          324818        69092.0       2480716.0
3  20210307    AS               0            NaN          2140.0
4  20210307    AZ          826454        56519.0       3073010.0

   Pending_Tests totalTestResultsSource  Total_Tests  Currently_Hospitalized  \
0            NaN        totalTestsViral      1731628                    33.0
1            NaN  totalTestsPeopleViral      2323788                   494.0
2            NaN        totalTestsViral      2736442                   335.0
3            NaN        totalTestsViral         2140                     NaN
4            NaN        totalTestsViral      7908105                   963.0

   Total_Hospitalized  …  dataQualityGrade  deathIncrease  \
0              1293.0  …              None              0
1             45976.0  …              None             -1
2             14926.0  …              None             22
3                 NaN  …              None              0
4             57907.0  …              None              5

   hospitalizedIncrease                                      hash  \
0                     0  dc4bccd4bb885349d7e94d6fed058e285d4be164
1                     0  997207b430824ea40b8eb8506c19a93e07bc972e
2                    11  50921aeefba3e30d31623aa495b47fb2ecc72fae
```

```
3                     0  f77912d0b80d579fbb6202fa1a90554fc4dc1443
4                    44  0437a7a96f4471666f775e63e86923eb5cbd8cdf

   commercialScore negativeRegularScore negativeScore positiveScore  score  \
0                0                    0             0             0      0
1                0                    0             0             0      0
2                0                    0             0             0      0
3                0                    0             0             0      0
4                0                    0             0             0      0

   grade
0
1
2
3
4

[5 rows x 56 columns]
```

[9]:

```python
# Converted Date Columns - Standardized date format for analysis
```

[10]:

```python
# Step 4. Remove duplicate rows (if any)
df.drop_duplicates(inplace=True)

# View cleaned dataset
print(df.head())
```

```
       date State  Positive_Cases  probableCases  Negative_Cases  \
0  20210307    AK           56886            NaN             NaN
1  20210307    AL          499819       107742.0       1931711.0
2  20210307    AR          324818        69092.0       2480716.0
3  20210307    AS               0            NaN          2140.0
4  20210307    AZ          826454        56519.0       3073010.0

   Pending_Tests totalTestResultsSource  Total_Tests  Currently_Hospitalized  \
0            NaN        totalTestsViral      1731628                    33.0
1            NaN  totalTestsPeopleViral      2323788                   494.0
2            NaN        totalTestsViral      2736442                   335.0
3            NaN        totalTestsViral         2140                     NaN
4            NaN        totalTestsViral      7908105                   963.0

   Total_Hospitalized  …  dataQualityGrade  deathIncrease  \
0              1293.0  …              None              0
1             45976.0  …              None             -1
2             14926.0  …              None             22
3                 NaN  …              None              0
4             57907.0  …              None              5
```

```
   hospitalizedIncrease                                      hash  \
0                     0  dc4bccd4bb885349d7e94d6fed058e285d4be164
1                     0  997207b430824ea40b8eb8506c19a93e07bc972e
2                    11  50921aeefba3e30d31623aa495b47fb2ecc72fae
3                     0  f77912d0b80d579fbb6202fa1a90554fc4dc1443
4                    44  0437a7a96f4471666f775e63e86923eb5cbd8cdf

   commercialScore negativeRegularScore negativeScore positiveScore  score  \
0                0                    0             0             0      0
1                0                    0             0             0      0
2                0                    0             0             0      0
3                0                    0             0             0      0
4                0                    0             0             0      0

   grade
0
1
2
3
4

[5 rows x 56 columns]
```

[11]:

```
# Removed Duplicates - Ensured unique records
```

[12]:

```
# Step 5. Handle missing values (replace NaN with 0 for numeric columns)
numeric_cols = df.select_dtypes(include=['number']).columns
df[numeric_cols] = df[numeric_cols].fillna(0)

# View cleaned dataset
print(df.head())
```

```
       date State  Positive_Cases  probableCases  Negative_Cases  \
0  20210307    AK           56886            0.0             0.0
1  20210307    AL          499819       107742.0       1931711.0
2  20210307    AR          324818        69092.0       2480716.0
3  20210307    AS               0            0.0          2140.0
4  20210307    AZ          826454        56519.0       3073010.0

   Pending_Tests totalTestResultsSource  Total_Tests  Currently_Hospitalized  \
0            0.0        totalTestsViral      1731628                    33.0
1            0.0  totalTestsPeopleViral      2323788                   494.0
2            0.0        totalTestsViral      2736442                   335.0
3            0.0        totalTestsViral         2140                     0.0
4            0.0        totalTestsViral      7908105                   963.0

   Total_Hospitalized  …  dataQualityGrade  deathIncrease  \
0              1293.0  …              None              0
```

```
1             45976.0  …              None             -1
2             14926.0  …              None             22
3                 0.0  …              None              0
4             57907.0  …              None              5

   hospitalizedIncrease                                      hash  \
0                     0  dc4bccd4bb885349d7e94d6fed058e285d4be164
1                     0  997207b430824ea40b8eb8506c19a93e07bc972e
2                    11  50921aeefba3e30d31623aa495b47fb2ecc72fae
3                     0  f77912d0b80d579fbb6202fa1a90554fc4dc1443
4                    44  0437a7a96f4471666f775e63e86923eb5cbd8cdf

   commercialScore negativeRegularScore negativeScore positiveScore  score  \
0                0                    0             0             0      0
1                0                    0             0             0      0
2                0                    0             0             0      0
3                0                    0             0             0      0
4                0                    0             0             0      0

   grade
0
1
2
3
4

[5 rows x 56 columns]
```

[13]:
```
# Handled Missing Values - Replaced NaN with 0 for numerical fields
```

[14]:
```
# •         1 paragraph of the ethical implications of data wrangling specific to your datasource and the steps you completed answering the following questions:
# o         What changes were made to the data?
# o         Are there any legal or regulatory guidelines for your data or project topic?
# o         What risks could be created based on the transformations done?
# o         Did you make any assumptions in cleaning/transforming the data?
# o         How was your data sourced / verified for credibility?
# o         Was your data acquired in an ethical way?
# o         How would you mitigate any of the ethical implications you have identified?
```

In wrangling the COVID Tracking Project data, several transformations were applied, including renaming columns for clarity, standardizing state abbreviations, converting date formats, removing duplicates, and replacing missing numerical values with zero.

Since COVID-19 data impacts public health policies, it is subject to regulatory guidelines such as

CDC reporting standards and HIPAA (if linked to personal health data).

A key ethical risk is data misrepresentation—treating missing values as zero may incorrectly imply no cases rather than unreported data, potentially leading to inaccurate conclusions.

This assumption was made to ensure consistency, but it may not always be valid. The dataset was sourced from a now-archived public API maintained by The COVID Tracking Project, which aggregated information from state health departments, but variations in state-level reporting could impact accuracy. While the data was ethically collected, potential biases or errors in state reporting must be considered.

To mitigate ethical concerns, transparent documentation of data transformations, acknowledgment of data limitations, and cross-referencing with authoritative sources (e.g., CDC, WHO) are essential to ensure accurate and responsible data use.