Chronic Illness Analysis 2.0

January 18, 2025

[1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
```

0.0.1 Distribution of Chronic Condition Category

[3]:
```
import pandas as pd

# Specify file path
file_path = r'\\AR-FS01\users\MWD\Work Data\Python Source Files\Resident␣
↪Vaccination Mastersheet.xlsx'

# Read all columns except 'Resident Name'
df = pd.read_excel(file_path, sheet_name='Master Sheet')

df=df.drop(columns=['Resident Name'])
```

[4]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 210 entries, 0 to 209
Data columns (total 83 columns):
 #   Column                                   Non-Null Count  Dtype
---  ------                                   --------------  -----
 0   Resident ID                              210 non-null    int64
 1   Gender at Birth                          210 non-null    object
 2   Room                                     210 non-null    object
 3   Floor                                    210 non-null    int64
 4   DOB                                      210 non-null    datetime64[ns]
 5   AGE                                      210 non-null    int64
 6   Total Antibiotics                        210 non-null    int64
 7   COVID Pandemic Vaccine #1 Date           183 non-null    datetime64[ns]
 8   COVID Pandemic Vaccine #2 Date           171 non-null    datetime64[ns]
 9   21 - 22 COVID Vaccination Status         210 non-null    object
 10  22 - 23 COVID Vaccination  Date          173 non-null    object
 11  22 - 23 COVID Vaccination Status         210 non-null    object
```

```
12  24-25 COVID Dose Date                       112 non-null    datetime64[ns]
13  24-25 COVID Vaccination Status              210 non-null    object
14  Total COVID Vaccines                        210 non-null    int64
15  22-23 Flu Vaccination Date                  146 non-null    datetime64[ns]
16  22-23 Flu Vaccination Status                206 non-null    object
17  24-25 Flu Vaccination Date                  110 non-null    datetime64[ns]
18  24-25 Flu Vaccination Status                210 non-null    object
19  Pneumococcal Vaccination Date               163 non-null    datetime64[ns]
20  Pneumococcal Vaccine Type                   161 non-null    object
21  Pneumoccal Vaccine Status                   210 non-null    object
22  RSV Vaccination Date                        33 non-null     datetime64[ns]
23  RSV  Vaccination Status                     210 non-null    object
24  Total Infections Since 2022                 209 non-null    float64
25  Diabetes                                    207 non-null    float64
26  Pre-Diabetes                                207 non-null    float64
27  Hypertension                                207 non-null    float64
28  Hyperlipidemia                              207 non-null    float64
29  Hyperthyroidism                             207 non-null    float64
30  MI                                          207 non-null    float64
31  Heart Disease                               207 non-null    float64
32  CKD                                         207 non-null    float64
33  Anemia                                      207 non-null    float64
34  Indwelling Device?                          207 non-null    float64
35  Osteoporosis                                207 non-null    float64
36  Obesity                                     207 non-null    float64
37  Constipation                                207 non-null    float64
38  Incontinence                                207 non-null    float64
39  OSA                                         207 non-null    float64
40  COPD                                        207 non-null    float64
41  Asthma                                      207 non-null    float64
42  PE                                          207 non-null    float64
43  Ineffective Adherrence                      207 non-null    float64
44  EPS                                         207 non-null    float64
45  Chronic Pain                                207 non-null    float64
46  Nicotine Dependence/ Toacco Use Disorder    207 non-null    float64
47  Alcohol Use Disorder                        207 non-null    float64
48  GERD                                        207 non-null    float64
49  Insomnia                                    207 non-null    float64
50  Hypothyroidism                              207 non-null    float64
51  Iron or Vitamin Deficiency                  207 non-null    float64
52  Seasonal Allergies                          207 non-null    float64
53  Stimulant Dependence                        207 non-null    float64
54  Venous Insufficiency                        207 non-null    float64
55  Peripheral Arterial Disease                 207 non-null    float64
56  Cannabis Use Disorder                       207 non-null    float64
57  Liver Disease                               207 non-null    float64
58  Gout                                        207 non-null    float64
59  Latent TB                                   207 non-null    float64
```

```
 60  UC / Gastritis                          207 non-null    float64
 61  Dementia                                207 non-null    float64
 62  Dry Eyes                                207 non-null    float64
 63  RA / OA Arthritis                       207 non-null    float64
 64  Migraines / Headahes                    207 non-null    float64
 65  Addisons Disease                        207 non-null    float64
 66  Total Chronic Conditions                210 non-null    int64
 67  Admit Date 1                            44 non-null     datetime64[ns]
 68  Diagnosis Admit 1                       44 non-null     object
 69  Days Admitted Admit 1                   43 non-null     float64
 70  Admit Date 2                            18 non-null     object
 71  Diagnosis Admit 2                       17 non-null     object
 72  Days Admitted Admit 2                   16 non-null     float64
 73  Admit Date 3                            7 non-null      datetime64[ns]
 74  Diagnosis Admit  3                      7 non-null      object
 75  Days Admitted Admit3                    7 non-null      float64
 76  Admit Date 4                            4 non-null      datetime64[ns]
 77  Diagnosis Admit  4                      4 non-null      object
 78  Days Admitted Admit 4                   4 non-null      float64
 79  Total Days Admitted                     210 non-null    int64
 80  Total Number of Admissions              210 non-null    int64
 81  Change in Weight                        210 non-null    object
 82  Weight change as percent                191 non-null    float64
dtypes: datetime64[ns](11), float64(47), int64(8), object(17)
memory usage: 136.3+ KB
```

[5]:
```
df.columns
```

[5]:
```
Index(['Resident ID', 'Gender at Birth', 'Room', 'Floor', 'DOB', 'AGE',
       'Total Antibiotics', 'COVID Pandemic Vaccine #1 Date',
       'COVID Pandemic Vaccine #2 Date', '21 - 22 COVID Vaccination Status',
       '22 - 23 COVID Vaccination  Date', '22 - 23 COVID Vaccination Status',
       '24-25 COVID Dose Date', '24-25 COVID Vaccination Status',
       'Total COVID Vaccines', '22-23 Flu Vaccination Date',
       '22-23 Flu Vaccination Status', '24-25 Flu Vaccination Date',
       '24-25 Flu Vaccination Status', 'Pneumococcal Vaccination Date',
       'Pneumococcal Vaccine Type', 'Pneumoccal Vaccine Status',
       'RSV Vaccination Date', 'RSV  Vaccination Status',
       'Total Infections Since 2022', 'Diabetes', 'Pre-Diabetes',
       'Hypertension', 'Hyperlipidemia', 'Hyperthyroidism', 'MI',
       'Heart Disease', 'CKD', 'Anemia', 'Indwelling Device?', 'Osteoporosis',
       'Obesity', 'Constipation', 'Incontinence', 'OSA', 'COPD', 'Asthma',
       'PE', 'Ineffective Adherrence', 'EPS', 'Chronic Pain',
       'Nicotine Dependence/ Toacco Use Disorder', 'Alcohol Use Disorder',
       'GERD', 'Insomnia', 'Hypothyroidism', 'Iron or Vitamin Deficiency',
       'Seasonal Allergies', 'Stimulant Dependence', 'Venous Insufficiency',
       'Peripheral Arterial Disease', 'Cannabis Use Disorder', 'Liver Disease',
```

```
       'Gout', 'Latent TB', 'UC / Gastritis', 'Dementia', 'Dry Eyes',
       'RA / OA Arthritis', 'Migraines / Headahes', 'Addisons Disease',
       'Total Chronic Conditions', 'Admit Date 1', 'Diagnosis Admit 1',
       'Days Admitted Admit 1', 'Admit Date 2', 'Diagnosis Admit 2',
       'Days Admitted Admit 2', 'Admit Date 3', 'Diagnosis Admit  3',
       'Days Admitted Admit3', 'Admit Date 4', 'Diagnosis Admit  4',
       'Days Admitted Admit 4', 'Total Days Admitted',
       'Total Number of Admissions', 'Change in Weight',
       'Weight change as percent'],
      dtype='object')
```

[6]:
```
df.head()
```

[6]:
```
   Resident ID Gender at Birth   Room  Floor        DOB  AGE  \
0            1          Female  325-2      3 1994-01-01   30
1            2          Female  229-1      2 1992-03-31   32
2            3            Male  426-2      4 1974-11-12   50
3            4          Female  301-1      3 1963-09-11   61
4          227            Male  306-1      3 1978-08-14   46

   Total Antibiotics COVID Pandemic Vaccine #1 Date  \
0                  0                     2021-12-15
1                  0                     2021-07-01
2                  0                     2021-01-03
3                  0                     2021-01-06
4                  0                            NaT

  COVID Pandemic Vaccine #2 Date 21 - 22 COVID Vaccination Status  …  \
0                     2022-01-24                       Up To Date  …
1                     2021-07-01                       Up To Date  …
2                     2021-01-31                       Up To Date  …
3                     2021-02-03                       Up To Date  …
4                            NaT                         Declined  …

  Admit Date 3 Diagnosis Admit  3 Days Admitted Admit3 Admit Date 4  \
0          NaT                NaN                  NaN          NaT
1          NaT                NaN                  NaN          NaT
2          NaT                NaN                  NaN          NaT
3          NaT                NaN                  NaN          NaT
4          NaT                NaN                  NaN          NaT

   Diagnosis Admit  4 Days Admitted Admit 4 Total Days Admitted  \
0                 NaN                   NaN                   3
1                 NaN                   NaN                   0
2                 NaN                   NaN                   0
3                 NaN                   NaN                   0
4                 NaN                   NaN                   0
```

```
  Total Number of Admissions Change in Weight Weight change as percent
0                          1             55.8                 0.200719
1                          0             -4.8                -0.017033
2                          0             -9.2                -0.051627
3                          0               -5                -0.022936
4                          0              7.5                 0.032895

[5 rows x 83 columns]
```

[7]:
```
# Initialize a flag in the global scope
if not hasattr(df, '_weight_transformed'):
    df['Weight change as percent'] = df['Weight change as percent'] * 100    #␣
 ↪Ensures that the expression is only carried out once
    df['Weight change as percent'] = df['Weight change as percent'].round(2)
    df._weight_transformed = True  # Set the flag to True
```

[8]:
```
df.head()
```

[8]:
```
   Resident ID Gender at Birth   Room  Floor         DOB  AGE  \
0            1          Female  325-2      3  1994-01-01   30
1            2          Female  229-1      2  1992-03-31   32
2            3            Male  426-2      4  1974-11-12   50
3            4          Female  301-1      3  1963-09-11   61
4          227            Male  306-1      3  1978-08-14   46

   Total Antibiotics COVID Pandemic Vaccine #1 Date  \
0                  0                     2021-12-15
1                  0                     2021-07-01
2                  0                     2021-01-03
3                  0                     2021-01-06
4                  0                            NaT

  COVID Pandemic Vaccine #2 Date 21 - 22 COVID Vaccination Status  …  \
0                     2022-01-24                       Up To Date  …
1                     2021-07-01                       Up To Date  …
2                     2021-01-31                       Up To Date  …
3                     2021-02-03                       Up To Date  …
4                            NaT                         Declined  …

  Admit Date 3 Diagnosis Admit  3 Days Admitted Admit3 Admit Date 4  \
0          NaT               NaN                    NaN          NaT
1          NaT               NaN                    NaN          NaT
2          NaT               NaN                    NaN          NaT
3          NaT               NaN                    NaN          NaT
4          NaT               NaN                    NaN          NaT
```

```
   Diagnosis Admit  4 Days Admitted Admit 4 Total Days Admitted  \
0                NaN                    NaN                     3
1                NaN                    NaN                     0
2                NaN                    NaN                     0
3                NaN                    NaN                     0
4                NaN                    NaN                     0

  Total Number of Admissions Change in Weight Weight change as percent
0                          1             55.8                    20.07
1                          0             -4.8                    -1.70
2                          0             -9.2                    -5.16
3                          0               -5                    -2.29
4                          0              7.5                     3.29

[5 rows x 83 columns]
```

[9]:
```python
# count of Data Types
df.dtypes.value_counts()
```

[9]:
```
float64           47
object            17
datetime64[ns]    11
int64              8
Name: count, dtype: int64
```

0.0.2 Create New Dataframe with only Numberical Data

[19]:
```python
df_numerical = df.drop(columns=[
    'Room', 'DOB',
    'COVID Pandemic Vaccine #1 Date', 'COVID Pandemic Vaccine #2 Date',
    '22 - 23 COVID Vaccination  Date', '24-25 COVID Dose Date',
    '22-23 Flu Vaccination Date', '24-25 Flu Vaccination Date',
    'Pneumococcal Vaccination Date', 'Pneumococcal Vaccine Type',
    'RSV Vaccination Date', 'Admit Date 1', 'Diagnosis Admit 1',
    'Admit Date 2', 'Diagnosis Admit 2', 'Admit Date 3', 'Diagnosis Admit  3',
    'Admit Date 4', 'Diagnosis Admit  4'
])

df_numerical.head()
```

[19]:
```
   Resident ID Gender at Birth  Floor  AGE  Total Antibiotics  \
0            1          Female      3   30                  0
1            2          Female      2   32                  0
2            3            Male      4   50                  0
3            4          Female      3   61                  0
4          227            Male      3   46                  0
```

```
  21 - 22 COVID Vaccination Status 22 - 23 COVID Vaccination Status  \
0                       Up To Date                       Vaccinated
1                       Up To Date                       Vaccinated
2                       Up To Date                       Vaccinated
3                       Up To Date                       Vaccinated
4                         Declined                   Not Vaccinated

  24-25 COVID Vaccination Status  Total COVID Vaccines  \
0                 Not Up To Date                     3
1                     Up To Date                     4
2                     Up To Date                     4
3                     Up To Date                     4
4                     Up To Date                     1

  22-23 Flu Vaccination Status  … Addisons Disease Total Chronic Conditions  \
0               Not Vaccinated  …              0.0                        7
1                     Declined  …              0.0                        9
2                   Vaccinated  …              0.0                        5
3                   Vaccinated  …              0.0                       16
4               Not Vaccinated  …              0.0                        3

  Days Admitted Admit 1  Days Admitted Admit 2  Days Admitted Admit3  \
0                   3.0                    NaN                   NaN
1                   NaN                    NaN                   NaN
2                   NaN                    NaN                   NaN
3                   NaN                    NaN                   NaN
4                   NaN                    NaN                   NaN

   Days Admitted Admit 4  Total Days Admitted  Total Number of Admissions  \
0                    NaN                    3                           1
1                    NaN                    0                           0
2                    NaN                    0                           0
3                    NaN                    0                           0
4                    NaN                    0                           0

   Change in Weight  Weight change as percent
0              55.8                     20.07
1              -4.8                     -1.70
2              -9.2                     -5.16
3                -5                     -2.29
4               7.5                      3.29

[5 rows x 64 columns]
```

0.0.3 Check for Missing Values

[22]:
```
# Check for nulls

df_numerical.isnull().sum()
```

[22]:
```
Resident ID                    0
Gender at Birth                0
Floor                          0
AGE                            0
Total Antibiotics              0
                              …
Days Admitted Admit 4        206
Total Days Admitted            0
Total Number of Admissions     0
Change in Weight               0
Weight change as percent      19
Length: 64, dtype: int64
```

[24]:
```
# Replacing null values with 0's in days Admitted Columns

admit_columns =['Days Admitted Admit 1', 'Days Admitted Admit 2', 'Days␣
↪Admitted Admit3', 'Days Admitted Admit 4']

df_numerical[admit_columns] = df_numerical[admit_columns].fillna(0)
```

[26]:
```
# Replacing NaN Values from the Percent weight change column to 0

df_numerical['Weight change as percent'] = df_numerical['Weight change as␣
↪percent'].fillna(0)
```

[28]:
```
# Replacing Null Values with 0's in Change in Weight

df_numerical['Change in Weight'] = df_numerical['Change in Weight'].fillna(0)
```

[30]:
```
# Replacing NaN Values in Days admitted Columns

    # Adresses Days Admitted 1
df_numerical['Days Admitted Admit 1'] = df_numerical['Days Admitted Admit 1'].
↪fillna(0)

    # Adresses Days Admitted 2
df_numerical['Days Admitted Admit 2'] = df_numerical['Days Admitted Admit 2'].
↪fillna(0)

    # Adresses Days Admitted 3
```

```
df_numerical['Days Admitted Admit3'] = df_numerical['Days Admitted Admit3'].
↪fillna(0)

    # Adresses Days Admitted 4
df_numerical['Days Admitted Admit 4'] = df_numerical['Days Admitted Admit 4'].
↪fillna(0)
```

[32]:
```
# Drop Null Values From Resident ID Column

df_numerical = df_numerical.dropna(subset=['Resident ID'])
```

0.1 Encoding Columns

Encoding Vaccine Status Columns

[40]:
```
# Encoding 21 - 22 COVID Vaccination Status Column
df_numerical['21 - 22 COVID Vaccination Status'] = df_numerical['21 - 22 COVID␣
↪Vaccination Status'].apply(
    lambda x: 1 if x == 'Vaccinated' else(0.5 if x == 'Partially Vaccinated'␣
↪else 0)    # Replaces Not Vaccinated w/ 0, Partially Vaccinated w/ 0.5,␣
↪Vaccinated w/ 1
)

# Encoding 22 - 23 COVID Vaccination Status Column
df_numerical['22 - 23 COVID Vaccination Status'] = df_numerical['22 - 23 COVID␣
↪Vaccination Status'].apply(
    lambda x: 1 if x == 'Vaccinated' else(0.5 if x == 'Partially Vaccinated'␣
↪else 0)    # Replaces Not Vaccinated w/ 0, Partially Vaccinated w/ 0.5,␣
↪Vaccinated w/ 1
)

# Encoding 24 - 25 COVID Vaccination Status Column
df_numerical['24-25 COVID Vaccination Status'] = df_numerical['24-25 COVID␣
↪Vaccination Status'].apply(
    lambda x: 1 if x == 'Vaccinated' else(0.5 if x == 'Partially Vaccinated'␣
↪else 0)    # Replaces Not Vaccinated w/ 0, Partially Vaccinated w/ 0.5,␣
↪Vaccinated w/ 1
)

# Encoding 22-23 Flu Vaccination Status Column
df_numerical['22-23 Flu Vaccination Status'] = df_numerical['22-23 Flu␣
↪Vaccination Status'].apply(
    lambda x: 1 if x == 'Vaccinated' else(0.5 if x == 'Partially Vaccinated'␣
↪else 0)    # Replaces Not Vaccinated w/ 0, Partially Vaccinated w/ 0.5,␣
↪Vaccinated w/ 1
)

# Encoding 24-25 Flu Vaccination Status Column
```

```
df_numerical['24-25 Flu Vaccination Status'] = df_numerical['24-25 Flu␣
↪Vaccination Status'].apply(
    lambda x: 1 if x == 'Vaccinated' else(0.5 if x == 'Partially Vaccinated'␣
↪else 0)    # Replaces Not Vaccinated w/ 0, Partially Vaccinated w/ 0.5,␣
↪Vaccinated w/ 1
)

# Encoding Pneumoccal Vaccine Status Column
df_numerical['Pneumoccal Vaccine Status'] = df_numerical['Pneumoccal Vaccine␣
↪Status'].apply(
    lambda x: 1 if x == 'Vaccinated' else(0.5 if x == 'Partially Vaccinated'␣
↪else 0)    # Replaces Not Vaccinated w/ 0, Partially Vaccinated w/ 0.5,␣
↪Vaccinated w/ 1
)

# Encodnig RSV Vaccination Status Column
df_numerical['RSV  Vaccination Status'] = df_numerical['RSV  Vaccination␣
↪Status'].apply(
    lambda x: 1 if x == 'Up To Date' else(0.5 if x == 'Not Eligible' else 0)   ␣
↪# Replaces Not Vaccinated w/ 0, Not Eligible w/ 0.5, Vaccinated w/ 1
)
```

Enconding Gender Column

[43]:
```
# Gender At Birth Column
df_numerical['Gender at Birth'] = df_numerical['Gender at Birth'].apply(
    lambda x: 1 if x == 'Male' else 0)
```

0.2 Adding Age Categories

[46]:
```
# Create an Age Category such as 18-30, 30-40 etc

def age_group(x):                   # Creates function that bins age into groups
    if 18 <= x <= 30:
        return '18-30'
    elif 31 <= x <= 40:
        return '31-40'
    elif 41 <= x <= 50:
        return '41-50'
    elif 51 <= x <= 60:
        return '51-60'
    elif 61 <= x <= 70:
        return '61-70'
    elif 71 <= x <= 80:
        return '71-80'
    elif 81 <= x <= 100:
        return '81-100'
```

```
    else:
        return 'Unknown'

# Apply new Age Group Column to ORIGINAL DATAFRAME
df['Age Group'] = df_numerical['AGE'].apply(age_group)        # Adds new Age␣
 ↪Group Column to Original DataFrame df

# Also add to df_numerical
df_numerical['Age Group'] = df_numerical['AGE'].apply(age_group)

df_numerical.head()
```

[46]:

```
   Resident ID  Gender at Birth  Floor  AGE  Total Antibiotics  \
0            1                0      3   30                  0
1            2                0      2   32                  0
2            3                1      4   50                  0
3            4                0      3   61                  0
4          227                1      3   46                  0

   21 - 22 COVID Vaccination Status  22 - 23 COVID Vaccination Status  \
0                               0.0                                 1
1                               0.0                                 1
2                               0.0                                 1
3                               0.0                                 1
4                               0.0                                 0

   24-25 COVID Vaccination Status  Total COVID Vaccines  \
0                               0                     3
1                               0                     4
2                               0                     4
3                               0                     4
4                               0                     1

   22-23 Flu Vaccination Status  …  Total Chronic Conditions  \
0                             0  …                         7
1                             0  …                         9
2                             1  …                         5
3                             1  …                        16
4                             0  …                         3

   Days Admitted Admit 1  Days Admitted Admit 2  Days Admitted Admit3  \
0                    3.0                    0.0                   0.0
1                    0.0                    0.0                   0.0
2                    0.0                    0.0                   0.0
3                    0.0                    0.0                   0.0
4                    0.0                    0.0                   0.0
```

```
   Days Admitted Admit 4  Total Days Admitted  Total Number of Admissions  \
0                    0.0                    3                           1
1                    0.0                    0                           0
2                    0.0                    0                           0
3                    0.0                    0                           0
4                    0.0                    0                           0

   Change in Weight  Weight change as percent  Age Group
0              55.8                     20.07      18-30
1              -4.8                     -1.70      31-40
2              -9.2                     -5.16      41-50
3                -5                     -2.29      61-70
4               7.5                      3.29      41-50

[5 rows x 65 columns]
```

[48]:
```python
# Encode Age Categories

def age_categories1 (x):
    if x == '18-30':
        return 0        # Assigns 18-30 a value of 0
    if x == '31-40':
        return 1        # Assigns 31-40 a value of 1
    if x == '41-50':
        return 2        # Assigns 41-50 a value of 2
    if x == '51-60':
        return 3        # Assigns 51-60 a value of 3
    if x == '61-70':
        return 4        # Assigns 61-70 a value of 4
    if x == '81-100':
        return 5        # Assigns 81-100 a value of 5
    else:
        return '6'      # Assigns all else a value of 6


# apply function to the Age Category Column of df_numerical
df_numerical['Age Group'] = df_numerical['Age Group'].apply(age_categories1)

df_numerical['Age Group'].head()
```

[48]:
```
0    0
1    1
2    2
3    4
4    2
Name: Age Group, dtype: object
```

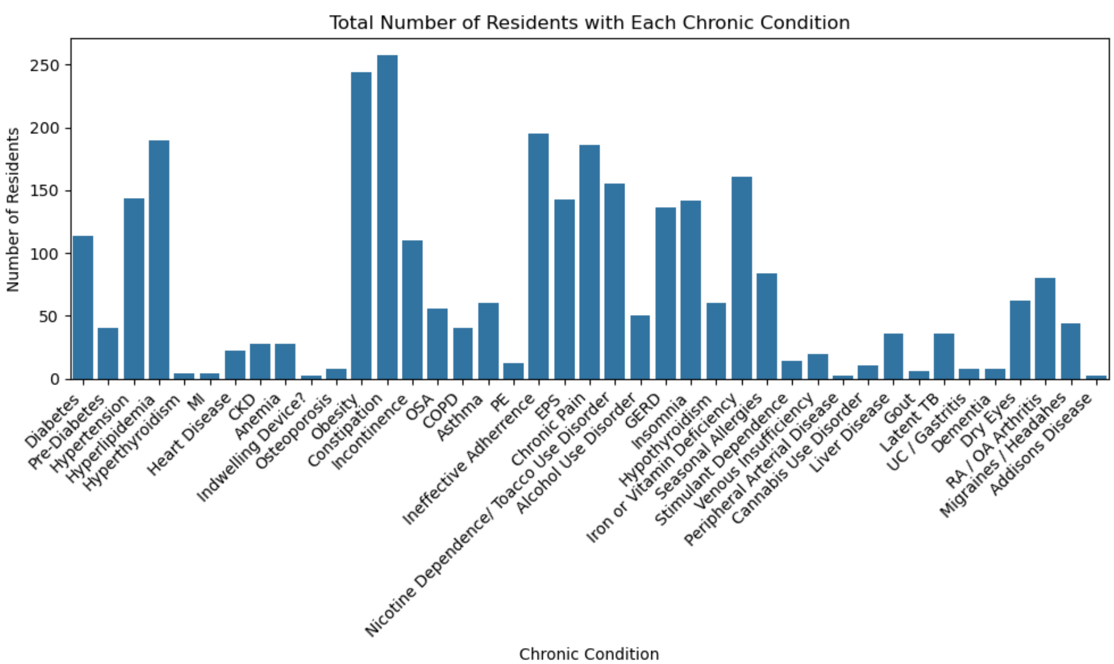
0.2.1 Total of Chronic Conditions

[51]:
```python
# Select columns from 'Diabetes' to 'Addisons Disease' by column names
selected_df = df_numerical.loc[:, 'Diabetes':'Addisons Disease']

# Add a new row at the bottom with the sum of each column
selected_df.loc['Total'] = selected_df.sum()

# Display the resulting DataFrame
selected_df.tail()
```

[51]:
```
       Diabetes  Pre-Diabetes  Hypertension  Hyperlipidemia  Hyperthyroidism  \
206         1.0           0.0           1.0             0.0              0.0
207         1.0           0.0           1.0             0.0              0.0
208         1.0           0.0           1.0             1.0              0.0
209         1.0           0.0           1.0             0.0              0.0
Total      57.0          20.0          72.0            95.0              2.0

        MI  Heart Disease   CKD  Anemia  Indwelling Device?  …  \
206    0.0            0.0   0.0     0.0                 0.0  …
207    0.0            0.0   0.0     1.0                 0.0  …
208    0.0            0.0   0.0     0.0                 0.0  …
209    0.0            0.0   0.0     0.0                 0.0  …
Total  2.0           11.0  14.0    14.0                 1.0  …

       Cannabis Use Disorder  Liver Disease  Gout  Latent TB  UC / Gastritis  \
206                      0.0            1.0   0.0        0.0             0.0
207                      0.0            0.0   0.0        1.0             0.0
208                      0.0            0.0   0.0        0.0             0.0
209                      0.0            0.0   1.0        0.0             0.0
Total                    6.0           18.0   3.0       18.0             4.0

       Dementia  Dry Eyes  RA / OA Arthritis  Migraines / Headahes  \
206         0.0       0.0                0.0                   0.0
207         0.0       1.0                0.0                   0.0
208         0.0       0.0                1.0                   0.0
209         0.0       0.0                1.0                   0.0
Total       4.0      31.0               40.0                  22.0

       Addisons Disease
206                 0.0
207                 0.0
208                 0.0
209                 0.0
Total               1.0

[5 rows x 41 columns]
```

0.2.2 Visualize Chronic Illness Totals

[54]:
```python
# Ensure there are no duplicate rows in selected_df
selected_df = selected_df.drop_duplicates()

# Replace blanks (NaN) with 0
selected_df = selected_df.fillna(0)

# Sum each chronic condition column to get accurate counts
chronic_condition_sums = selected_df.sum()

# Plot the corrected sums
plt.figure(figsize=(10, 6))
sns.barplot(x=chronic_condition_sums.index, y=chronic_condition_sums.values)

# Set chart title and labels
plt.title('Total Number of Residents with Each Chronic Condition')
plt.xlabel('Chronic Condition')
plt.ylabel('Number of Residents')

# Rotate the x-axis labels for better readability
plt.xticks(rotation=45, ha='right')

# Show the plot
plt.tight_layout()
plt.show()
```



0.3 Standardizing Data

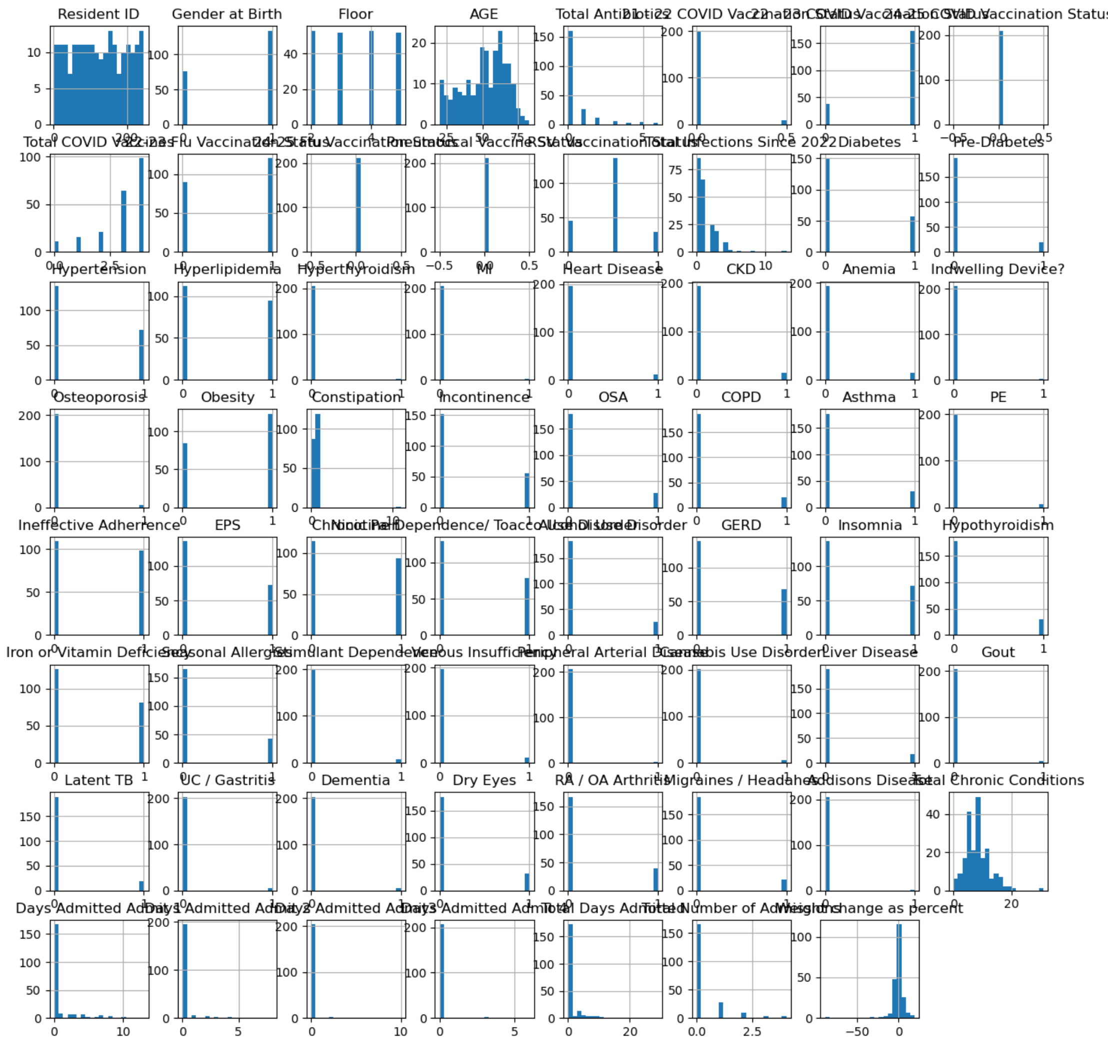
[56]:
```
# View ranges of columns (min vs max) to determine which fields need scaling
df_numerical.describe()
```

[56]:
```
       Resident ID  Gender at Birth       Floor         AGE  \
count   210.000000       210.000000  210.000000  210.000000
mean    124.295238         0.638095    3.495238   51.300000
std      72.059309         0.481700    1.120695   15.162296
min       1.000000         0.000000    2.000000   20.000000
25%      64.500000         0.000000    2.250000   39.000000
50%     123.000000         1.000000    3.500000   53.000000
75%     187.500000         1.000000    4.000000   63.000000
max     246.000000         1.000000    5.000000   83.000000

       Total Antibiotics  21 - 22 COVID Vaccination Status  \
count         210.000000                        210.000000
mean            0.490476                          0.021429
std             1.133401                          0.101509
min             0.000000                          0.000000
25%             0.000000                          0.000000
50%             0.000000                          0.000000
75%             0.000000                          0.000000
max             6.000000                          0.500000

       22 - 23 COVID Vaccination Status  24-25 COVID Vaccination Status  \
count                        210.000000                           210.0
mean                           0.823810                             0.0
std                            0.381892                             0.0
min                            0.000000                             0.0
25%                            1.000000                             0.0
50%                            1.000000                             0.0
75%                            1.000000                             0.0
max                            1.000000                             0.0

       Total COVID Vaccines  22-23 Flu Vaccination Status  …  \
count            210.000000                    210.000000  …
mean               3.057143                      0.571429  …
std                1.160173                      0.496054  …
min                0.000000                      0.000000  …
25%                3.000000                      0.000000  …
50%                3.000000                      1.000000  …
75%                4.000000                      1.000000  …
max                4.000000                      1.000000  …

       Migraines / Headahes  Addisons Disease  Total Chronic Conditions  \
count            207.000000        207.000000                210.000000
```

```
mean                 0.106280           0.004831                     8.347619
std                  0.308943           0.069505                     4.156982
min                  0.000000           0.000000                     0.000000
25%                  0.000000           0.000000                     6.000000
50%                  0.000000           0.000000                     8.000000
75%                  0.000000           0.000000                    10.000000
max                  1.000000           1.000000                    31.000000

       Days Admitted Admit 1  Days Admitted Admit 2  Days Admitted Admit3  \
count             210.000000             210.000000            210.000000
mean                0.909524               0.219048              0.142857
std                 2.282952               0.943180              0.922121
min                 0.000000               0.000000              0.000000
25%                 0.000000               0.000000              0.000000
50%                 0.000000               0.000000              0.000000
75%                 0.000000               0.000000              0.000000
max                13.000000               8.000000             10.000000

       Days Admitted Admit 4  Total Days Admitted  Total Number of Admissions  \
count             210.000000           210.000000                  210.000000
mean                0.071429             1.342857                    0.347619
std                 0.544343             3.685815                    0.799593
min                 0.000000             0.000000                    0.000000
25%                 0.000000             0.000000                    0.000000
50%                 0.000000             0.000000                    0.000000
75%                 0.000000             0.000000                    0.000000
max                 6.000000            29.000000                    4.000000

       Weight change as percent
count                210.000000
mean                  -0.256190
std                    8.202048
min                  -89.590000
25%                   -2.102500
50%                    0.000000
75%                    2.597500
max                   20.070000

[8 rows x 63 columns]
```

0.3.1 Data Distribution

[60]:
```python
# Plot histograms to view distributions
df_numerical.hist(bins=20, figsize=(15, 15))
plt.show()
```



[61]: `df_numerical.head()`

```
[61]:    Resident ID  Gender at Birth  Floor  AGE  Total Antibiotics  \
      0            1                0      3   30                  0
      1            2                0      2   32                  0
      2            3                1      4   50                  0
      3            4                0      3   61                  0
      4          227                1      3   46                  0

         21 - 22 COVID Vaccination Status  22 - 23 COVID Vaccination Status  \
      0                               0.0                                 1
      1                               0.0                                 1
      2                               0.0                                 1
      3                               0.0                                 1
      4                               0.0                                 0
```

```
   24-25 COVID Vaccination Status  Total COVID Vaccines  \
0                               0                     3
1                               0                     4
2                               0                     4
3                               0                     4
4                               0                     1

   22-23 Flu Vaccination Status  …  Total Chronic Conditions  \
0                             0  …                         7
1                             0  …                         9
2                             1  …                         5
3                             1  …                        16
4                             0  …                         3

   Days Admitted Admit 1  Days Admitted Admit 2  Days Admitted Admit3  \
0                    3.0                    0.0                   0.0
1                    0.0                    0.0                   0.0
2                    0.0                    0.0                   0.0
3                    0.0                    0.0                   0.0
4                    0.0                    0.0                   0.0

   Days Admitted Admit 4  Total Days Admitted  Total Number of Admissions  \
0                    0.0                    3                           1
1                    0.0                    0                           0
2                    0.0                    0                           0
3                    0.0                    0                           0
4                    0.0                    0                           0

   Change in Weight  Weight change as percent  Age Group
0              55.8                     20.07          0
1              -4.8                     -1.70          1
2              -9.2                     -5.16          2
3                -5                     -2.29          4
4               7.5                      3.29          2

[5 rows x 65 columns]
```

0.4 Scaling

Begin Scaling Columns

- ***Separate Continuous and Binary Columns:*** Identify continuous columns and apply standard scaling only to those.
  - ***Continuous Columns***: contain numerical data that can take on any value within a range
  - ***Binary Columns***: Contain data with only 2 distinct values
- ***Apply Standard Scaling:*** Scale columns with higher ranges (like AGE, Total Chronic

Conditions, and Total Days Admitted) to have a mean of 0 and a standard deviation of 1.

Fix New Non-numeric Columns

[65]:
```
# Change in Weight

# Check for Non-Numeric values
print(df_numerical['Change in Weight'].unique())
```

```
[55.80000000000001 -4.800000000000011 -9.199999999999989 -5 7.5
 -64.39999999999998 10 4 -6.800000000000011 -0.4000000000000057
 4.599999999999966 -8.400000000000006 19.80000000000001 1
 -11.400000000000006 ' ' 0.8999999999999773 1.1999999999999886 -3
 9.199999999999989 -1 4.800000000000011 -1.4000000000000057
 16.599999999999994 9 3.3000000000000114 3.799999999999983
 0.5999999999999943 -7.599999999999994 0 -14 -3.3999999999999773
 -0.5999999999999943 -5.400000000000006 2.6000000000000227 -144.6
 -6.799999999999983 -14.400000000000006 5 -0.799999999999983
 3.5999999999999943 -2 4.400000000000006 6.400000000000006 -9 6
 -7.600000000000023 9.599999999999994 -5.399999999999977
 -1.8000000000000114 1.5999999999999943 31.599999999999994 -17
 -9.399999999999977 -2.8000000000000114 -7.400000000000006 -11
 -3.4000000000000057 -6.200000000000017 7 -18.399999999999977 -6
 -1.3000000000000114 -5.799999999999983 23.599999999999994
 13.599999999999994 -8.800000000000011 8 3.3999999999999773
 -15.799999999999983 27.900000000000006 9.800000000000011
 2.3999999999999773 8.400000000000006 2.8000000000000114 8.600000000000023
 5.400000000000006 -0.8000000000000114 3.4000000000000057
 -12.200000000000017 -2.5999999999999943 -30 -5.200000000000003
 29.400000000000034 7.799999999999983 2.200000000000003
 -31.600000000000023 -3.1999999999999886 -0.9000000000000057
 6.399999999999977 -19.5 16.5 2.5999999999999943 -39.599999999999994
 -0.19999999999998863 12 14.800000000000011 21 -2.200000000000017
 15.199999999999989 5.200000000000017 -2.6999999999999886
 -6.400000000000006 -15.800000000000011 7.800000000000011 18
 39.80000000000001 7.600000000000023 3.6000000000000227 4.399999999999977
 4.200000000000017 10.800000000000011 12.199999999999989
 -27.399999999999977 11.599999999999994 11.199999999999989
 10.200000000000017 20.19999999999999 -1.1999999999999886
 -1.5999999999999943 -2.1999999999999886 -3.8000000000000114
 12.399999999999977 5.800000000000011 -55.599999999999994
 0.8000000000000114 -13.399999999999977 -8.099999999999994
 -6.300000000000011 -4.300000000000011 -7.800000000000011
 1.3999999999999773 0.20000000000001705 1.6000000000000227
 -4.299999999999983 9.449999999999989 5.199999999999989 17.400000000000006
 7.399999999999977 -16 2 -4.099999999999994 -20.399999999999977]
```

[66]:
```
# Replace Blank String with NaN
```

```
df_numerical['Change in Weight'] = df_numerical['Change in Weight'].replace('␣
↪', np.nan)

# Convert to Numeric
df_numerical['Change in Weight'] = pd.to_numeric(df_numerical['Change in␣
↪Weight'], errors='coerce')
```

```
C:\Users\mwd\AppData\Local\Temp\ipykernel_3508\174136052.py:2: FutureWarning:
Downcasting behavior in `replace` is deprecated and will be removed in a future
version. To retain the old behavior, explicitly call
`result.infer_objects(copy=False)`. To opt-in to the future behavior, set
`pd.set_option('future.no_silent_downcasting', True)`
  df_numerical['Change in Weight'] = df_numerical['Change in Weight'].replace('
', np.nan)
```

[67]:

```
# Age Group

# Check for non numeric values
df_numerical['Age Group'].unique()
```

[67]:

```
array([0, 1, 2, 4, 3, '6', 5], dtype=object)
```

[73]:

```
#  Change Age Group to Int
df_numerical['Age Group'] = df_numerical['Age Group'].astype(int)
```

0.4.1 Retain 1 non Scaled Dataframe (df_numerical) and one scaled Dataframe (df_scaled)

[76]:

```
df_numerical.dtypes
```

[76]:

```
Resident ID                      int64
Gender at Birth                  int64
Floor                            int64
AGE                              int64
Total Antibiotics                int64
                                ...
Total Days Admitted              int64
Total Number of Admissions       int64
Change in Weight               float64
Weight change as percent       float64
Age Group                        int32
Length: 65, dtype: object
```

[78]:

```
# Define Columns to Scale
continuous_columns = ['AGE','Total Antibiotics', 'Total COVID Vaccines', 'Total␣
↪Infections Since 2022',
                      'Total Chronic Conditions', 'Total Number of Admissions',
                      'Total Days Admitted', 'Days Admitted Admit 1',
```

```
                    'Days Admitted Admit 2', 'Days Admitted Admit3', 'Days␣
↪Admitted Admit 4', 'Change in Weight', 'Weight change as percent']


# Create a Copy of df_numerical to apply scaling too
df_scaled = df_numerical.copy()

# Initialize Scaler
scaler_infections = StandardScaler()

# Apply scaling to Continuous Columns in df_scaled
df_scaled[continuous_columns] = scaler_infections.
↪fit_transform(df_scaled[continuous_columns])

# Now df_numerical remains unchanged, and df_scaled contains the scaled values
df_scaled
```

[78]:

```
     Resident ID  Gender at Birth  Floor       AGE  Total Antibiotics  \
0              1                0      3 -1.408157          -0.433781
1              2                0      2 -1.275936          -0.433781
2              3                1      4 -0.085944          -0.433781
3              4                0      3  0.641273          -0.433781
4            227                1      3 -0.350387          -0.433781
..           …              …    …       …                …
205          204                1      5  0.707384          -0.433781
206          205                1      5 -1.342047          -0.433781
207          206                1      3  0.905716           3.988261
208          207                1      4  1.104048          -0.433781
209          208                0      3  0.046277          -0.433781

     21 - 22 COVID Vaccination Status  22 - 23 COVID Vaccination Status  \
0                                 0.0                                 1
1                                 0.0                                 1
2                                 0.0                                 1
3                                 0.0                                 1
4                                 0.0                                 0
..                                …                               …
205                               0.0                                 1
206                               0.0                                 1
207                               0.0                                 1
208                               0.0                                 1
209                               0.5                                 1

     24-25 COVID Vaccination Status  Total COVID Vaccines  \
0                                 0             -0.049371
1                                 0              0.814629
2                                 0              0.814629
```

```
3                                0            0.814629
4                                0           -1.777371
..                             …                …
205                              0           -0.049371
206                              0           -0.049371
207                              0            0.814629
208                              0            0.814629
209                              0           -0.913371

     22-23 Flu Vaccination Status  …  Total Chronic Conditions  \
0                               0  …                 -0.324957
1                               0  …                  0.157311
2                               1  …                 -0.807225
3                               1  …                  1.845249
4                               0  …                 -1.289493
..                            …  …                      …
205                             0  …                 -0.324957
206                             1  …                  0.157311
207                             1  …                  1.845249
208                             1  …                  0.639579
209                             0  …                  0.639579

     Days Admitted Admit 1  Days Admitted Admit 2  Days Admitted Admit3  \
0                 0.917878              -0.232799             -0.155292
1                -0.399350              -0.232799             -0.155292
2                -0.399350              -0.232799             -0.155292
3                -0.399350              -0.232799             -0.155292
4                -0.399350              -0.232799             -0.155292
..                     …                    …                    …
205              -0.399350              -0.232799             -0.155292
206              -0.399350              -0.232799             -0.155292
207               3.113258              -0.232799             -0.155292
208              -0.399350              -0.232799             -0.155292
209              -0.399350              -0.232799             -0.155292

     Days Admitted Admit 4  Total Days Admitted  Total Number of Admissions  \
0                -0.131533             0.450674                    0.817841
1                -0.131533            -0.365202                   -0.435784
2                -0.131533            -0.365202                   -0.435784
3                -0.131533            -0.365202                   -0.435784
4                -0.131533            -0.365202                   -0.435784
..                     …                  …                          …
205              -0.131533            -0.365202                   -0.435784
206              -0.131533            -0.365202                   -0.435784
207              -0.131533             1.810468                    0.817841
208              -0.131533            -0.365202                   -0.435784
209              -0.131533            -0.365202                   -0.435784
```

```
     Change in Weight  Weight change as percent  Age Group
0            3.446519                  2.484106          0
1           -0.256260                 -0.176451          1
2           -0.525109                 -0.599305          2
3           -0.268481                 -0.248556          4
4            0.495294                  0.433387          2
..                ...                       ...        ...
205               NaN                  0.031310          4
206         -0.213489                 -0.205782          1
207          0.012588                  0.005645          4
208          0.305878                  0.264735          4
209         -1.209451                 -0.912168          3

[210 rows x 65 columns]
```

[80]:
```
df_scaled.dtypes
```

[80]:
```
Resident ID                     int64
Gender at Birth                 int64
Floor                           int64
AGE                           float64
Total Antibiotics             float64
                                ...
Total Days Admitted           float64
Total Number of Admissions    float64
Change in Weight              float64
Weight change as percent      float64
Age Group                       int32
Length: 65, dtype: object
```

[82]:
```
# Retain a dataframe that is not scaled
df_numerical.head()
```

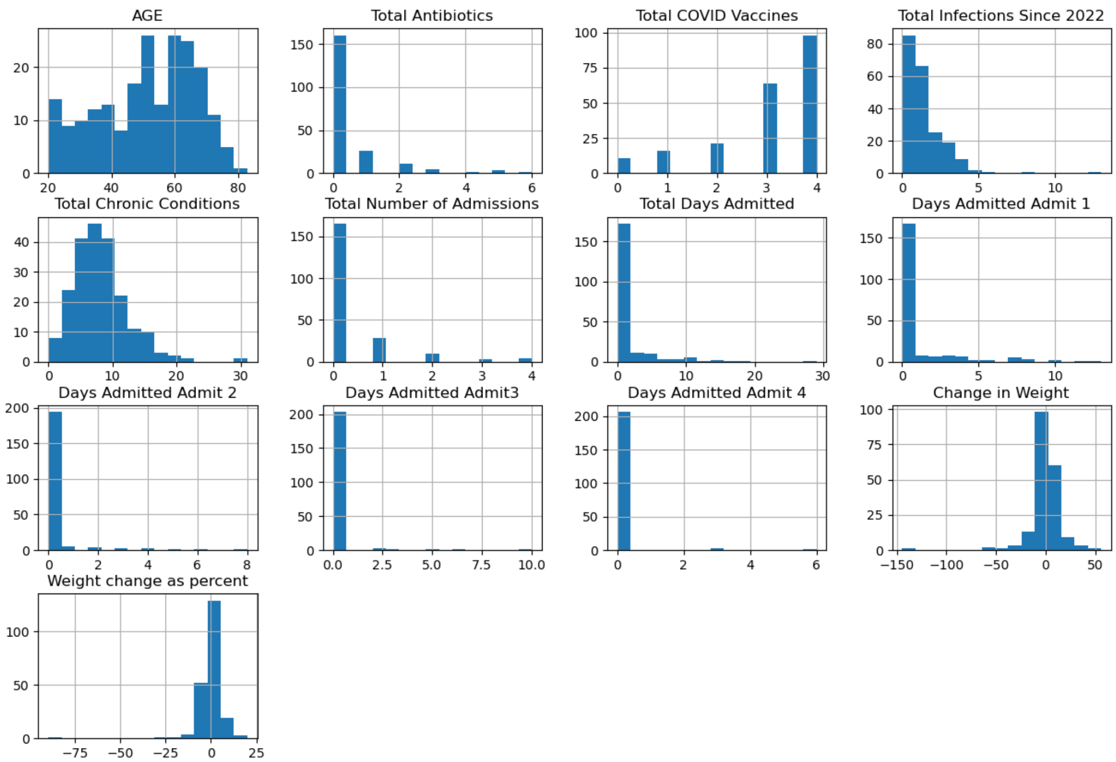
[82]:
```
   Resident ID  Gender at Birth  Floor  AGE  Total Antibiotics  \
0            1                0      3   30                  0
1            2                0      2   32                  0
2            3                1      4   50                  0
3            4                0      3   61                  0
4          227                1      3   46                  0

   21 - 22 COVID Vaccination Status  22 - 23 COVID Vaccination Status  \
0                               0.0                                 1
1                               0.0                                 1
2                               0.0                                 1
3                               0.0                                 1
4                               0.0                                 0
```

```
   24-25 COVID Vaccination Status  Total COVID Vaccines  \
0                              0                     3
1                              0                     4
2                              0                     4
3                              0                     4
4                              0                     1

   22-23 Flu Vaccination Status  …  Total Chronic Conditions  \
0                            0  …                         7
1                            0  …                         9
2                            1  …                         5
3                            1  …                        16
4                            0  …                         3

   Days Admitted Admit 1  Days Admitted Admit 2  Days Admitted Admit3  \
0                    3.0                    0.0                   0.0
1                    0.0                    0.0                   0.0
2                    0.0                    0.0                   0.0
3                    0.0                    0.0                   0.0
4                    0.0                    0.0                   0.0

   Days Admitted Admit 4  Total Days Admitted  Total Number of Admissions  \
0                    0.0                    3                           1
1                    0.0                    0                           0
2                    0.0                    0                           0
3                    0.0                    0                           0
4                    0.0                    0                           0

   Change in Weight  Weight change as percent  Age Group
0              55.8                     20.07          0
1              -4.8                     -1.70          1
2              -9.2                     -5.16          2
3              -5.0                     -2.29          4
4               7.5                      3.29          2

[5 rows x 65 columns]
```

0.5 Exploratory Data Analysis

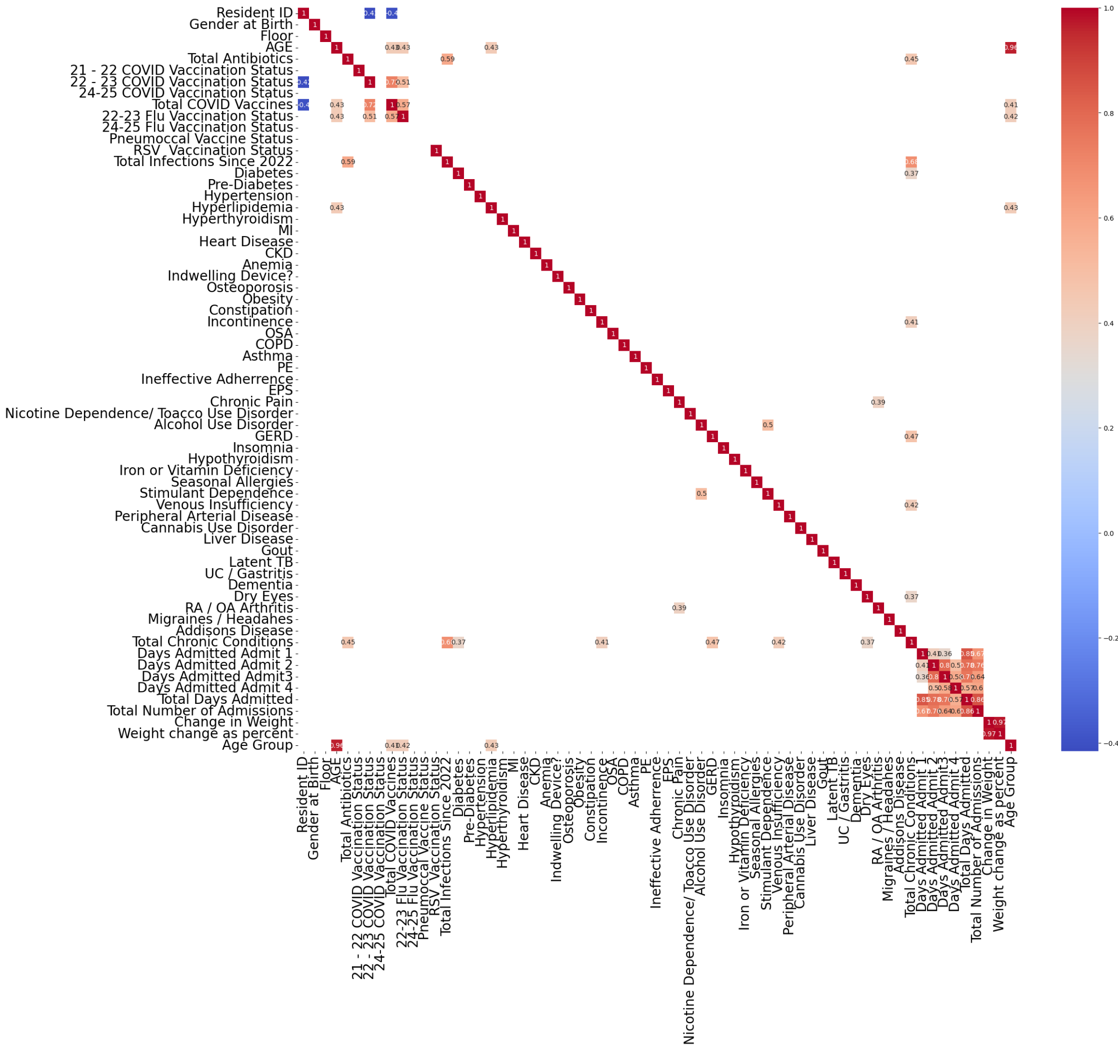
[85]:

```
# Histogram for continuous variables
df_numerical[continuous_columns].hist(bins=15, figsize=(15, 10))
plt.show()
```



0.5.1 Correlation Matrix

[88]:
```python
# Correlation Heatmap

corr_matrix = df_scaled.corr()

# Create a mask for correlations less than 0.5
mask = np.abs(corr_matrix) < 0.35

plt.figure(figsize=(24, 20))
sns.heatmap(corr_matrix, annot=True, cmap="coolwarm", mask=mask)
plt.xticks(rotation=90, fontsize = 20)
plt.yticks(rotation=0, fontsize = 20)
plt.show()
```



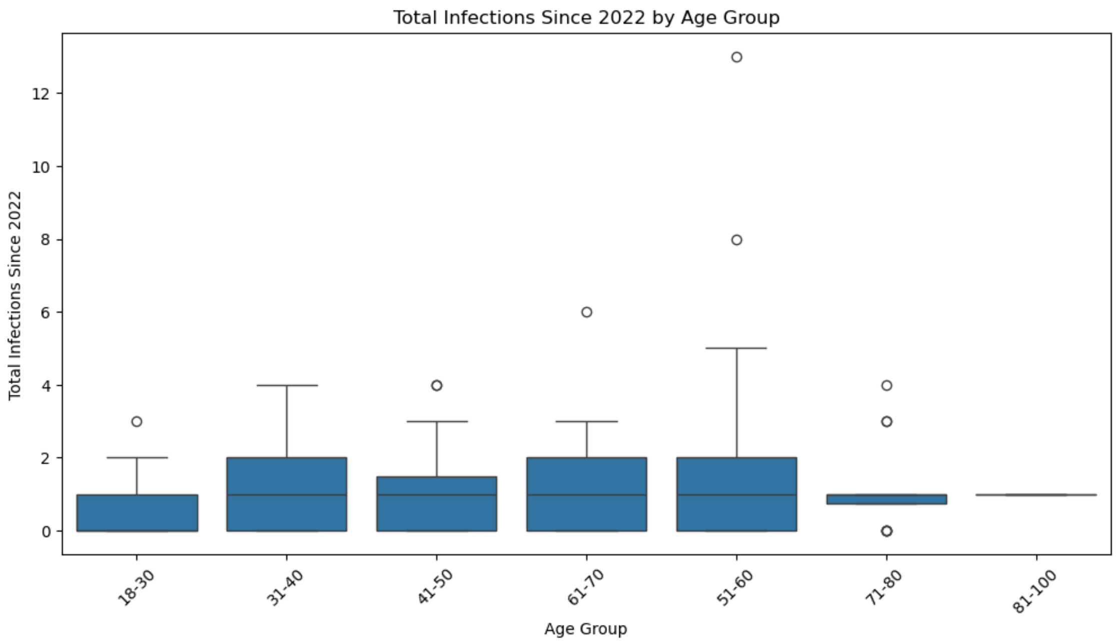
0.5.2 Box Plots

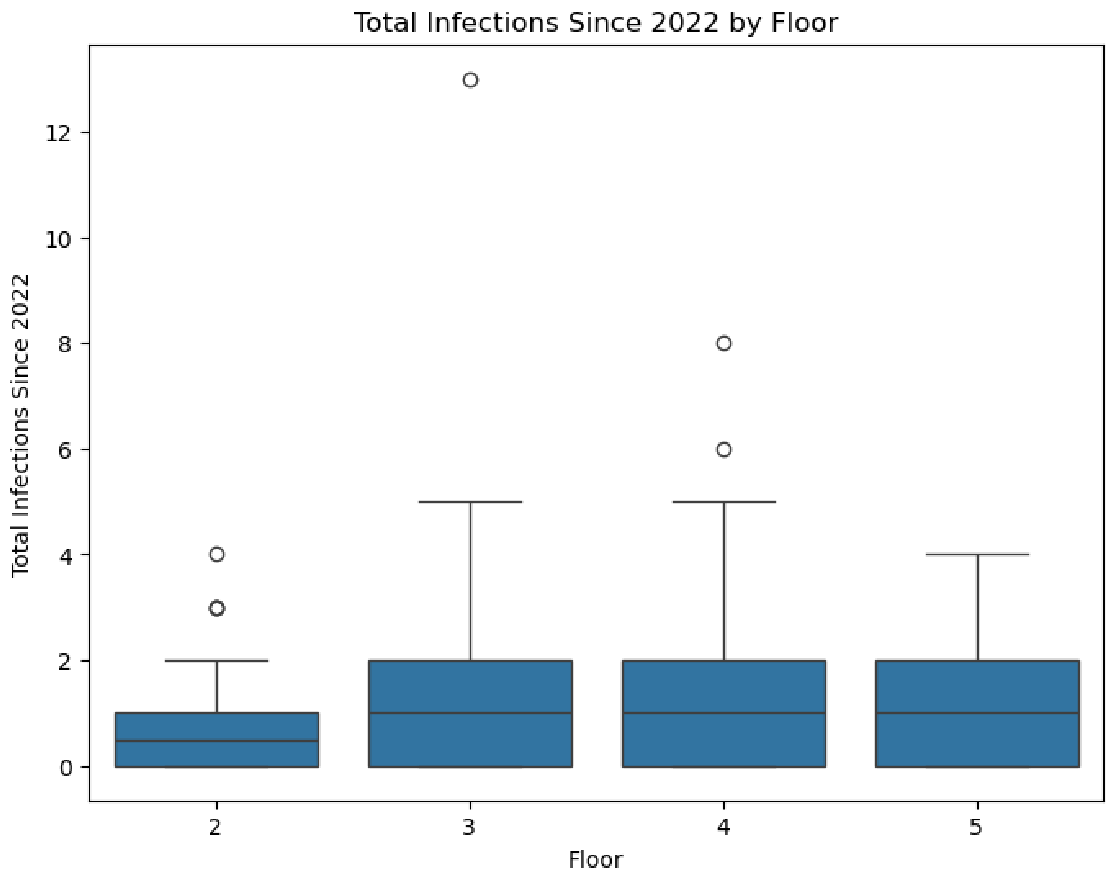
- Visualizing Infection Related Outcomes Across Categorical variables

Total Infections Since 2022 by Age Group

[92]:
```python
# Boxplot for Total Infections Since 2022 by Age Group
plt.figure(figsize=(12, 6))
sns.boxplot(x='Age Group', y='Total Infections Since 2022', data=df)
plt.title('Total Infections Since 2022 by Age Group')
plt.xlabel('Age Group')
plt.ylabel('Total Infections Since 2022')
plt.xticks(rotation=45)  # Rotate x-axis labels if needed for readability
plt.show()

# Boxplot for Total Infections Since 2022 by Floor
```

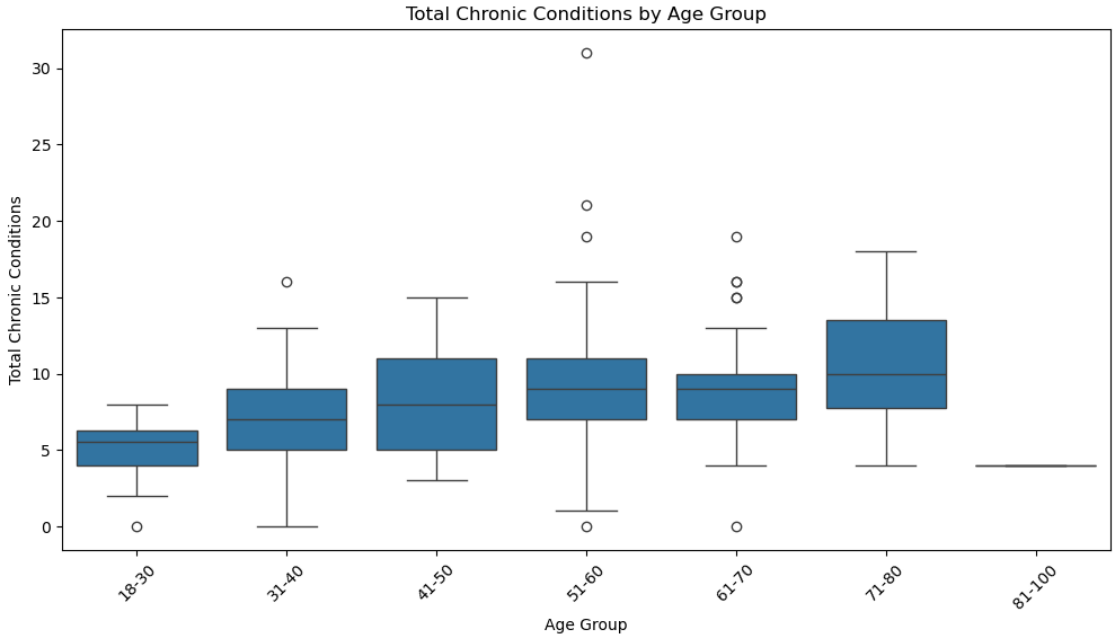
```
plt.figure(figsize=(8, 6))
sns.boxplot(x='Floor', y='Total Infections Since 2022', data=df)
plt.title('Total Infections Since 2022 by Floor')
plt.xlabel('Floor')
plt.ylabel('Total Infections Since 2022')
plt.show()
```



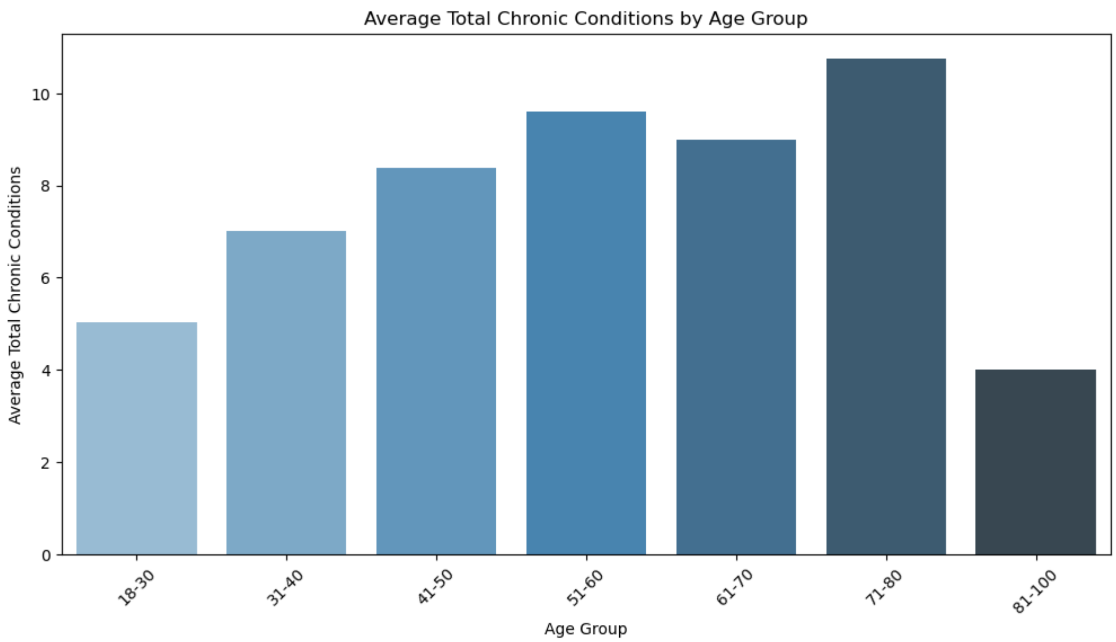


0.5.3 Age Group by Total Chronic Conditions

[95]:
```
# Setting 'Age Group' as a categorical variable with a specific order
age_group_order = ['18-30', '31-40', '41-50', '51-60', '61-70', '71-80',␣
 ↪'81-100']
df['Age Group'] = pd.Categorical(df['Age Group'], categories=age_group_order,␣
 ↪ordered=True)

# Create the boxplot, now ordered by Age Group
plt.figure(figsize=(12, 6))
sns.boxplot(x='Age Group', y='Total Chronic Conditions', data=df,␣
 ↪order=age_group_order)
plt.title('Total Chronic Conditions by Age Group')
plt.xlabel('Age Group')
plt.ylabel('Total Chronic Conditions')
plt.xticks(rotation=45)  # Rotate x-axis labels if needed
plt.show()
```



[97]:
```
# Total Chronic conditions by Age Group BAR CHART

# Assuming df is your DataFrame
# Set 'Age Group' as a categorical variable with a specific order
age_group_order = ['18-30', '31-40', '41-50', '51-60', '61-70', '71-80',␣
 ↪'81-100']
df['Age Group'] = pd.Categorical(df['Age Group'], categories=age_group_order,␣
 ↪ordered=True)

# Calculate the mean (or median) of 'Total Chronic Conditions' for each Age␣
 ↪Group
grouped_data = df.groupby('Age Group', observed=False)['Total Chronic␣
 ↪Conditions'].mean().reindex(age_group_order)

# Create the bar chart
plt.figure(figsize=(12, 6))
sns.barplot(x=grouped_data.index, y=grouped_data.values, palette="Blues_d",␣
 ↪hue=grouped_data.index, dodge=False, legend=False)
plt.title('Average Total Chronic Conditions by Age Group')
plt.xlabel('Age Group')
plt.ylabel('Average Total Chronic Conditions')
plt.xticks(rotation=45)  # Rotate x-axis labels if needed
plt.show()
```



0.5.4 Total Chronic Conditions by Total COVID Vaccines

- Using unscaled df (df_numerical)

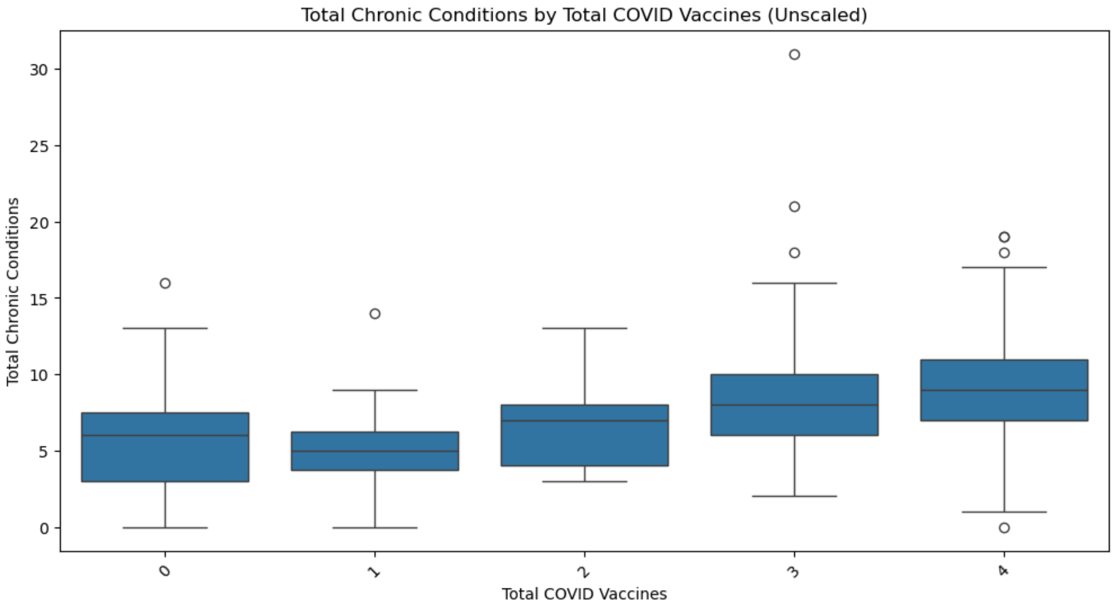
[100]:
```
# Group by 'Total COVID Vaccines' and calculate the mean of 'Total Chronic␣
↪Conditions'
average_by_vaccines = df_numerical.groupby('Total COVID Vaccines')['Total␣
↪Chronic Conditions'].mean()

# Display the results
for vaccine, avg_conditions in average_by_vaccines.items():
    print(f'{vaccine} COVID Vaccines, {round(avg_conditions, 0)} Average Total␣
↪Chronic Conditons')
```

```
0 COVID Vaccines, 6.0 Average Total Chronic Conditons
1 COVID Vaccines, 5.0 Average Total Chronic Conditons
2 COVID Vaccines, 7.0 Average Total Chronic Conditons
3 COVID Vaccines, 8.0 Average Total Chronic Conditons
4 COVID Vaccines, 9.0 Average Total Chronic Conditons
```

Total COVID Vaccines by Total Chronic Conditions

[103]:
```
# Plot with the original (unscaled) values from df_numerical
plt.figure(figsize=(12, 6))
sns.boxplot(x='Total COVID Vaccines', y='Total Chronic Conditions',␣
↪data=df_numerical)
plt.title('Total Chronic Conditions by Total COVID Vaccines (Unscaled)')
```

```
plt.xlabel('Total COVID Vaccines')
plt.ylabel('Total Chronic Conditions')
plt.xticks(rotation=45)  # Rotate labels for readability if needed
plt.show()
```



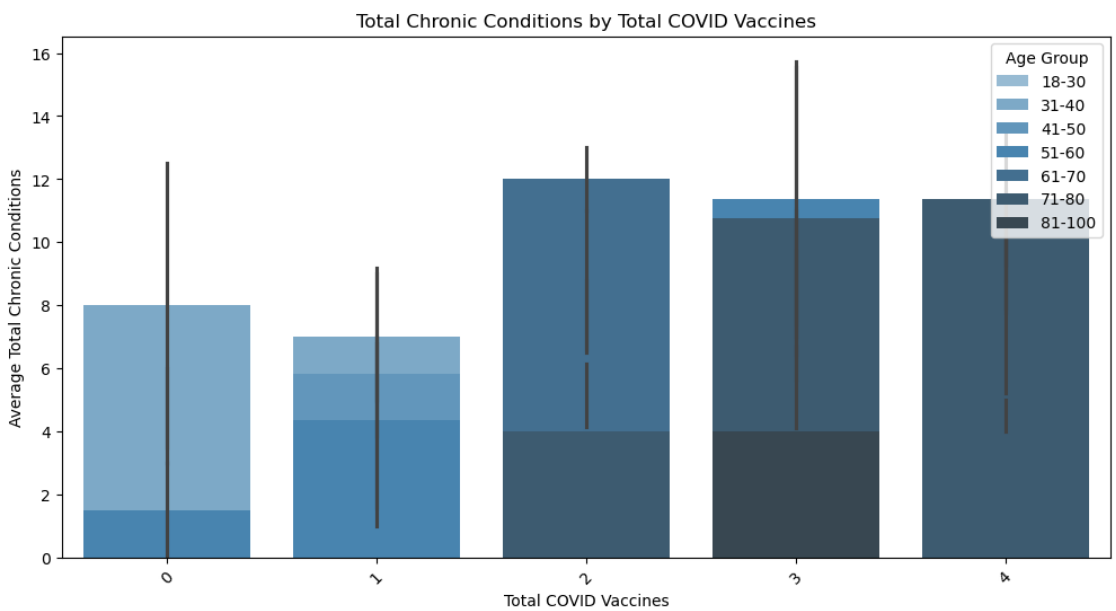
0.5.5 Total Chronic Conditions By Total Vaccines Filtered by Age Group

[106]:
```
## Group by 'Age Group' and calculate the total or mean of 'Total COVID␣
↪Vaccines'
grouped_vaccines_by_age = df.groupby('Age Group', observed = True)['Total COVID␣
↪Vaccines'].mean().reset_index()

# Display the results
print(grouped_vaccines_by_age)

f'Average Total Chronic Conditions {round(df_numerical['Total Chronic␣
↪Conditions'].mean(),0)}'
```

```
  Age Group  Total COVID Vaccines
0     18-30              2.142857
1     31-40              2.566667
2     41-50              2.870968
3     51-60              3.235294
4     61-70              3.584906
5     71-80              3.625000
6    81-100              3.000000
```

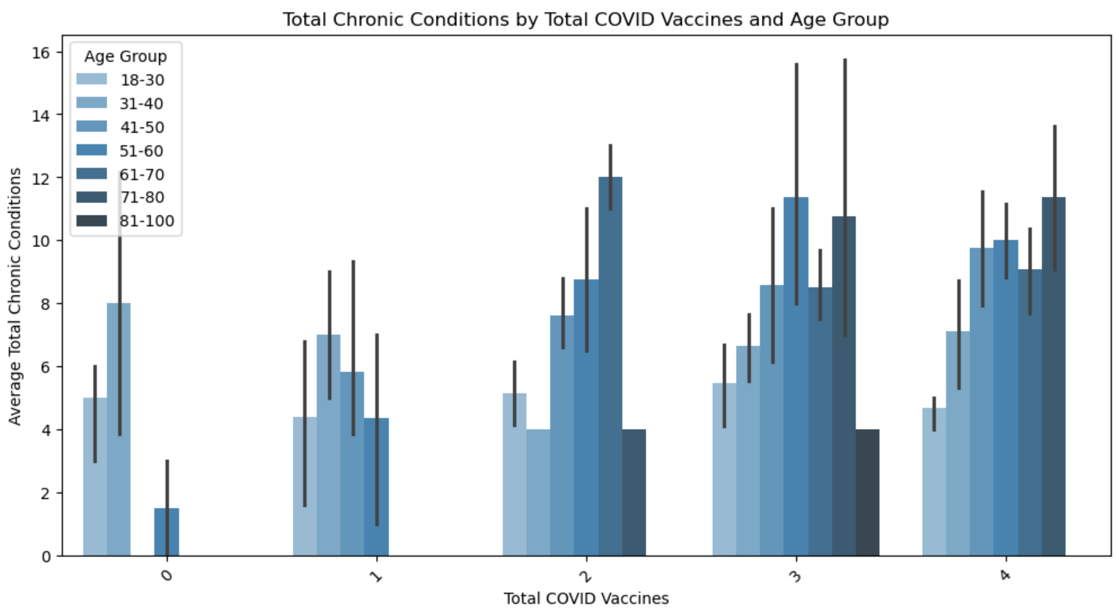
[106]: 'Average Total Chronic Conditions 8.0'

[108]:
```
# Total Chronic conditions by Total Chronic Conditions Grouped by Age Group

# Create the bar chart
plt.figure(figsize=(12, 6))

sns.barplot(
    x='Total COVID Vaccines',
    y='Total Chronic Conditions',
    hue='Age Group',  # Explicitly use 'Age Group' for grouping (if applicable)
    data=df,
    palette="Blues_d",
    dodge=False  # Use dodge=False for single-grouped bars
)

plt.title('Total Chronic Conditions by Total COVID Vaccines')
plt.xlabel('Total COVID Vaccines')
plt.ylabel('Average Total Chronic Conditions')
plt.xticks(rotation=45)
plt.legend(title='Age Group', loc='upper right')  # Optional, you can remove␣
↪this
plt.show()
```



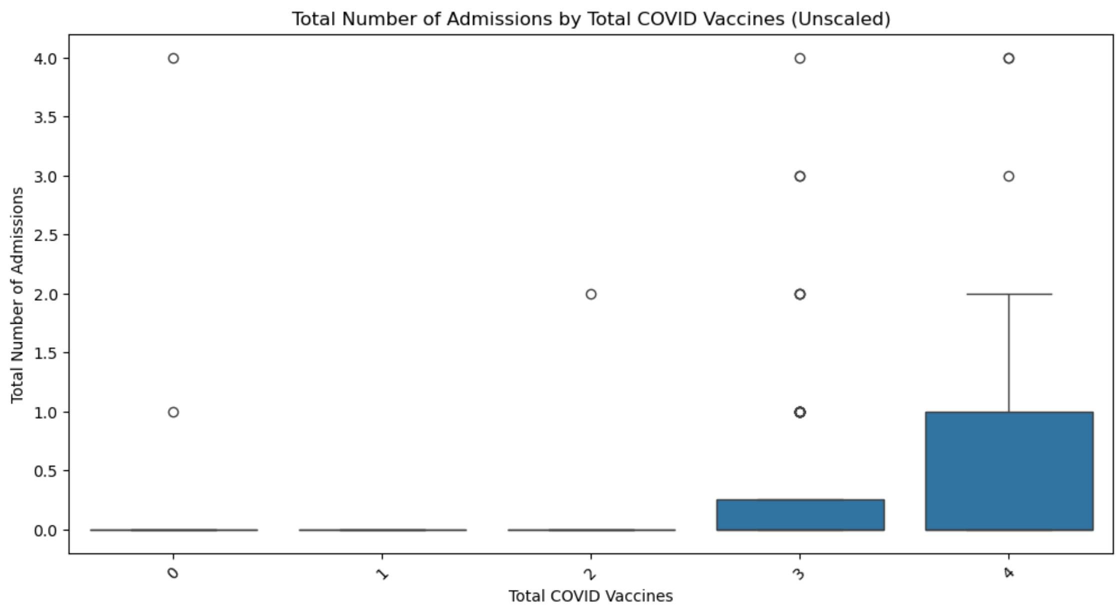
[109]:
```
# Total Chronic Conditions by Total COVID Vaccines Grouped by Age Group

plt.figure(figsize=(12, 6))
```

```
sns.barplot(
    x='Total COVID Vaccines',
    y='Total Chronic Conditions',
    hue='Age Group',
    data=df,
    palette="Blues_d"
)
plt.title('Total Chronic Conditions by Total COVID Vaccines and Age Group')
plt.xlabel('Total COVID Vaccines')
plt.ylabel('Average Total Chronic Conditions')
plt.xticks(rotation=45)
plt.legend(title='Age Group')
plt.show()
```



0.5.6 Total Number of Admissions by Total COVID Vaccines

[112]:
```
# Plot with the original values
plt.figure(figsize=(12, 6))
sns.boxplot(x='Total COVID Vaccines', y='Total Number of Admissions',␣
↪data=df_numerical)
plt.title('Total Number of Admissions by Total COVID Vaccines (Unscaled)')
plt.xlabel('Total COVID Vaccines')
plt.ylabel('Total Number of Admissions')
plt.xticks(rotation=45)  # Rotate x-axis labels if needed for readability
plt.show()
```



[114]:
```python
# Filter the DataFrame to include only rows where Total COVID Vaccines equals 4
filtered_df = df_numerical[df_numerical['Total COVID Vaccines'] == 4]

# Calculate the average age
average_age = filtered_df['AGE'].mean()

# Print the result
print(f"The average age of people with 4 Total COVID Vaccines is {average_age:.
 ↪2f} years.")
```

```
The average age of people with 4 Total COVID Vaccines is 56.67 years.
```

0.5.7 Age by Total Number of Admissions

Ordinal Encoding for Age Group

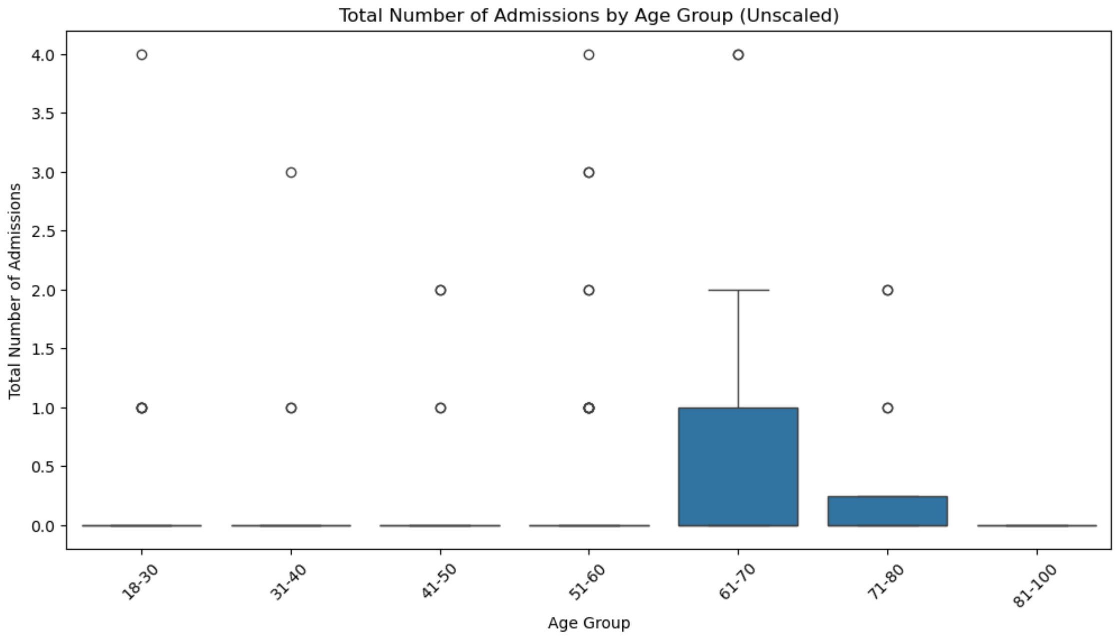
[119]:
```python
df_numerical.head()
```

[119]:
```
   Resident ID  Gender at Birth  Floor  AGE  Total Antibiotics  \
0            1                0      3   30                  0
1            2                0      2   32                  0
2            3                1      4   50                  0
3            4                0      3   61                  0
4          227                1      3   46                  0

   21 - 22 COVID Vaccination Status  22 - 23 COVID Vaccination Status  \
0                               0.0                                 1
1                               0.0                                 1
2                               0.0                                 1
```

```
3                              0.0                                  1
4                              0.0                                  0

   24-25 COVID Vaccination Status  Total COVID Vaccines  \
0                               0                     3
1                               0                     4
2                               0                     4
3                               0                     4
4                               0                     1

   22-23 Flu Vaccination Status  …  Total Chronic Conditions  \
0                             0  …                         7
1                             0  …                         9
2                             1  …                         5
3                             1  …                        16
4                             0  …                         3

   Days Admitted Admit 1  Days Admitted Admit 2  Days Admitted Admit3  \
0                    3.0                    0.0                   0.0
1                    0.0                    0.0                   0.0
2                    0.0                    0.0                   0.0
3                    0.0                    0.0                   0.0
4                    0.0                    0.0                   0.0

   Days Admitted Admit 4  Total Days Admitted  Total Number of Admissions  \
0                    0.0                    3                           1
1                    0.0                    0                           0
2                    0.0                    0                           0
3                    0.0                    0                           0
4                    0.0                    0                           0

   Change in Weight  Weight change as percent  Age Group
0              55.8                     20.07          0
1              -4.8                     -1.70          1
2              -9.2                     -5.16          2
3              -5.0                     -2.29          4
4               7.5                      3.29          2

[5 rows x 65 columns]
```

[121]:
```
# Plot with the original values
plt.figure(figsize=(12, 6))
sns.boxplot(x='Age Group', y='Total Number of Admissions', data=df)
plt.title('Total Number of Admissions by Age Group (Unscaled)')
plt.xlabel('Age Group')
plt.ylabel('Total Number of Admissions')
plt.xticks(rotation=45)  # Rotate x-axis labels if needed for readability
```

```
plt.show()
```



Not satisfied with the number of correlations and their strength. Therefore determined that additional feature creation was necessary to supplement the dataset and increase correlation scores and overall strengt, ultimately giving the prediction model a higher liklihood of performing at its optimal level.

[124]:

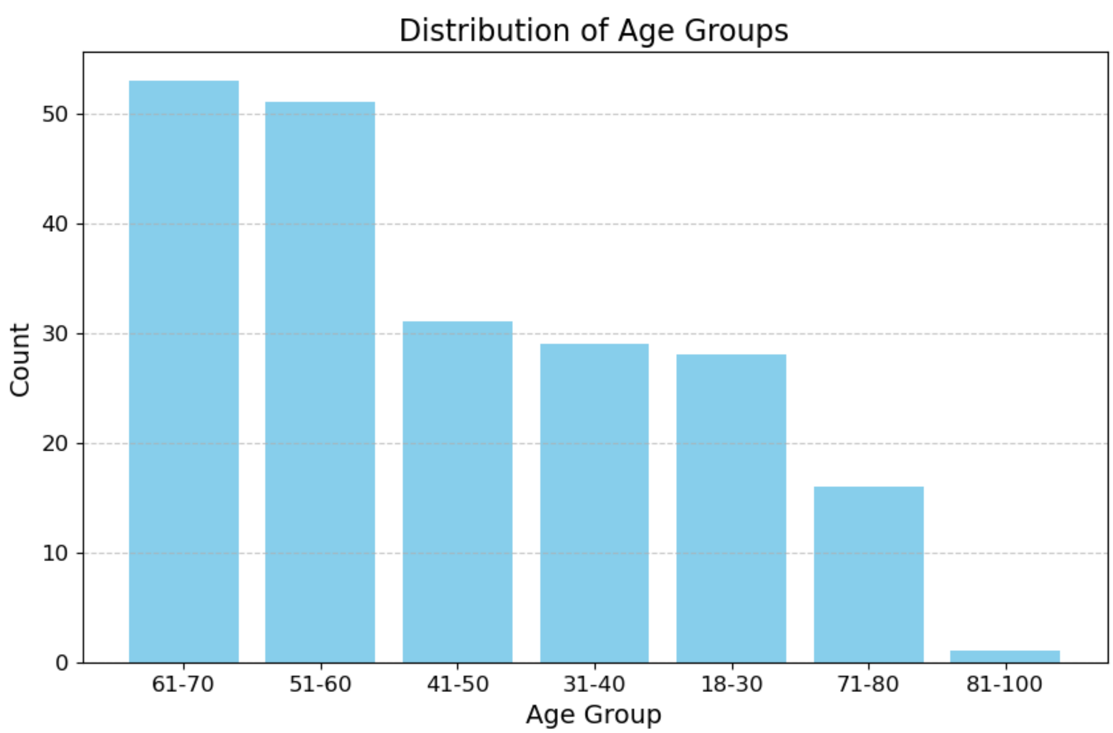
```
# Count occurrences of each age group
age_group_counts = df['Age Group'].value_counts()

# Display the counts
print(age_group_counts)
```

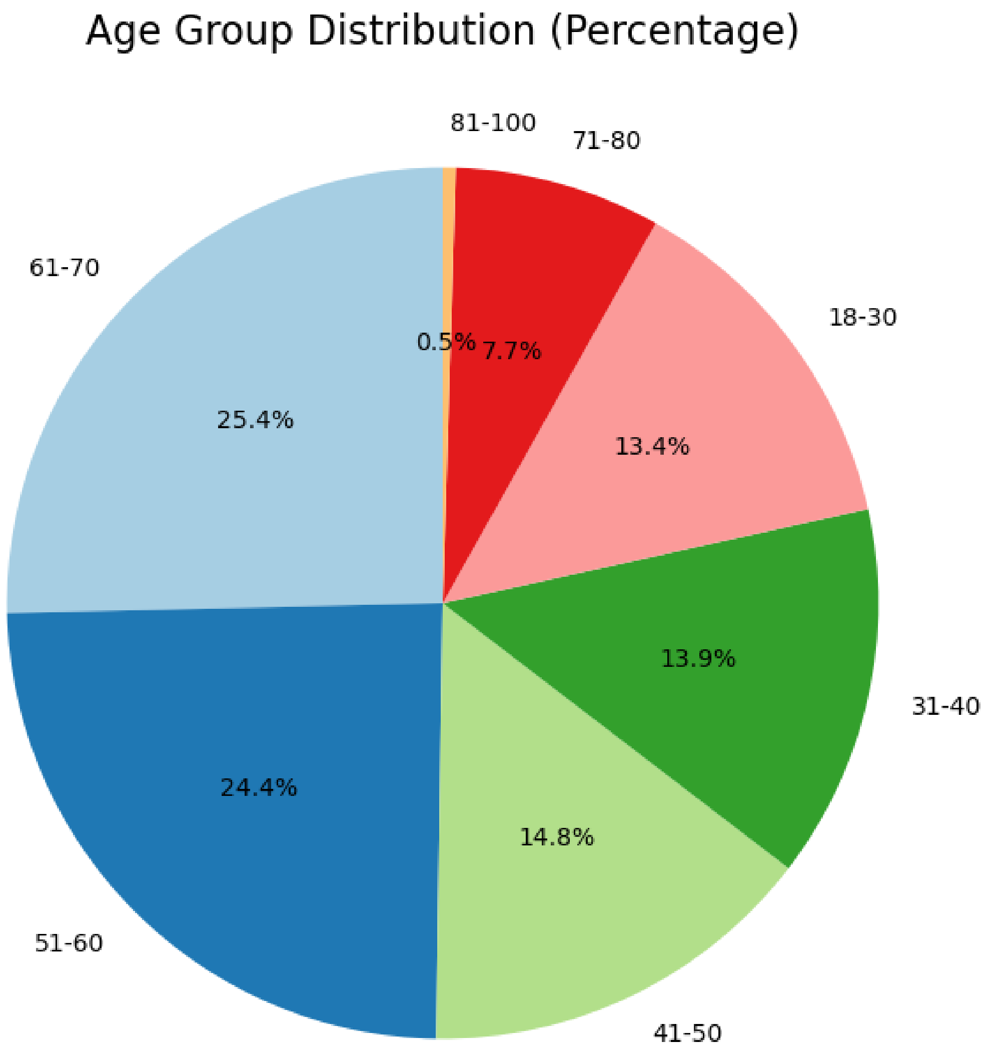
```
Age Group
61-70     53
51-60     51
41-50     31
31-40     30
18-30     28
71-80     16
81-100     1
Name: count, dtype: int64
```

[126]:

```
# Data for Age Groups and their counts
age_groups = ['61-70', '51-60', '41-50', '31-40', '18-30', '71-80', '81-100']
counts = [53, 51, 31, 29, 28, 16, 1]
```

```python
# Create a bar chart
plt.figure(figsize=(10, 6))
plt.bar(age_groups, counts, color='skyblue')
plt.title('Distribution of Age Groups', fontsize=16)
plt.xlabel('Age Group', fontsize=14)
plt.ylabel('Count', fontsize=14)
plt.xticks(fontsize=12)
plt.yticks(fontsize=12)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()

# Create a pie chart
plt.figure(figsize=(8, 8))
plt.pie(counts, labels=age_groups, autopct='%1.1f%%', startangle=90, colors=plt.
 ↪cm.Paired.colors)
plt.title('Age Group Distribution (Percentage)', fontsize=16)
plt.show()
```





1 Feature Engineering

- ***Combine Features*** Create new Features based on Existing Data
  - Health Risk Score:
  - Vaccine History Scire
  - Chronic Illness Categorization

1.1 Engineered Feature Health Risk Score

[130]:
```
# Assign weights to features for Health Risk Score

    # Assign column to df_numerical
df_numerical['Health Risk Score'] = (
    (df_numerical['Total Chronic Conditions'] * 2) +  # Higher weight for␣
 ↪chronic conditions
    (df_numerical['AGE'] * 0.5) +                    # Moderate weight for age
    (df_numerical['Diabetes'] * 1.5) +               # Specific weight for␣
 ↪diabetes
    (df_numerical['COPD'] * 1.5) +                   # COPD as a risk factor
    (df_numerical['Heart Disease'] * 2) +            # Higher risk weight for␣
 ↪heart disease
    (df_numerical['Weight change as percent'] * 1)   # Weight changes also␣
 ↪contribute
)

# Displaying the new column
print(df_numerical[['Health Risk Score']].head())


    # Assign column to df

df['Health Risk Score'] = (
    (df_numerical['Total Chronic Conditions'] * 2) +  # Higher weight for␣
 ↪chronic conditions
    (df_numerical['AGE'] * 0.5) +                    # Moderate weight for age
    (df_numerical['Diabetes'] * 1.5) +               # Specific weight for␣
 ↪diabetes
    (df_numerical['COPD'] * 1.5) +                   # COPD as a risk factor
    (df_numerical['Heart Disease'] * 2) +            # Higher risk weight for␣
 ↪heart disease
    (df_numerical['Weight change as percent'] * 1)   # Weight changes also␣
 ↪contribute
)

# Displaying the new column
print(df[['Health Risk Score']].head())
```

```
   Health Risk Score
0              49.07
1              32.30
2              29.84
3              63.21
4              32.29
   Health Risk Score
```

```
0               49.07
1               32.30
2               29.84
3               63.21
4               32.29
```

1.1.1 Distribution of Health Risk Scores

1.2 Engineered Feature: Vaccine History Score

- Assign a score of 1 for each fully vaccinated statuses
- Assign partial credit for partially vaccinated statuses

[134]:
```
# Calculating Vaccine History Score
df_numerical['Vaccine History Score'] = (
    df_numerical['21 - 22 COVID Vaccination Status'] +  # Max Score of 3 for␣
 ↪COVID Vaccines
    df_numerical['22 - 23 COVID Vaccination Status'] +  # Max Score of 3 for␣
 ↪COVID Vaccines
    df_numerical['24-25 COVID Vaccination Status'] +    # Max Score of 3 for␣
 ↪COVID Vaccines
    df_numerical['22-23 Flu Vaccination Status'] +      # Max Score of 2 for␣
 ↪Flu Vaccines
    df_numerical['24-25 Flu Vaccination Status'] +      # Max Score of 2 for␣
 ↪Flu Vaccines
    df_numerical['Pneumoccal Vaccine Status'] +         # Max Score of 1 for␣
 ↪Pneumococcal
    df_numerical['RSV  Vaccination Status']             # MAx Score of 1 for RSV
)
            # # Max Total Score of 7

# Displaying the new column
print(df_numerical[['Vaccine History Score']].head())
```

```
   Vaccine History Score
0                    1.5
1                    1.5
2                    2.5
3                    2.0
4                    0.5
```

1.3 Engineered Feature Categorize Chronic Conditions

- Separate Total Chronic Conditions into Low, Medium and High Risk
  - ***Low***: Less than 5 Chronic Conditions
  - ***Medium***: Between 5 and 10 Chronic Conditions
  - ***High***: More than 10 Chronic Conditions

[137]:
```python
# Categorize Chronic Conditions
def categorize_chronic_conditions(total_conditions):
    if total_conditions< 5:
        return 'Low Risk'
    elif 5<= total_conditions <= 10:
        return 'Medium Risk'
    else:
        return 'High Risk'

# Apply categorization to the 'Total Chronic Conditions' Column
df_numerical['Chronic Condition Category'] = df_numerical['Total Chronic␣
↪Conditions'].apply(categorize_chronic_conditions)

df_numerical[['Total Chronic Conditions', 'Chronic Condition Category']].head()
```

[137]:
```
   Total Chronic Conditions Chronic Condition Category
0                         7                Medium Risk
1                         9                Medium Risk
2                         5                Medium Risk
3                        16                  High Risk
4                         3                   Low Risk
```

1.3.1 Encode Chronic Conditions Category

- Low Risk = 1
- Medium Ris = 2
- High Risk = 3

[140]:
```python
df_numerical['Chronic Conditions Encoded'] = df_numerical['Chronic Condition␣
↪Category'].apply(
    lambda x: 1 if x == 'Low Risk' else(2 if x == 'Medium Risk' else 3)     #␣
↪Replaces Low Risk w/ 0, Medium Risk w/ 1, and High Risk w/ 2
)

df_numerical.drop(columns=['Chronic Conditions Encoded'])
```

[140]:
```
     Resident ID  Gender at Birth  Floor  AGE  Total Antibiotics  \
0              1                0      3   30                  0
1              2                0      2   32                  0
2              3                1      4   50                  0
3              4                0      3   61                  0
4            227                1      3   46                  0
..           …              …    …   …                …
205          204                1      5   62                  0
206          205                1      5   31                  0
207          206                1      3   65                  5
208          207                1      4   68                  0
```

```
209          208                0      3   52                 0

     21 - 22 COVID Vaccination Status  22 - 23 COVID Vaccination Status  \
0                                 0.0                                 1
1                                 0.0                                 1
2                                 0.0                                 1
3                                 0.0                                 1
4                                 0.0                                 0
..                                …                                …
205                               0.0                                 1
206                               0.0                                 1
207                               0.0                                 1
208                               0.0                                 1
209                               0.5                                 1

     24-25 COVID Vaccination Status  Total COVID Vaccines  \
0                                 0                     3
1                                 0                     4
2                                 0                     4
3                                 0                     4
4                                 0                     1
..                              …                   …
205                               0                     3
206                               0                     3
207                               0                     4
208                               0                     4
209                               0                     2

     22-23 Flu Vaccination Status  …  Days Admitted Admit3  \
0                               0  …                   0.0
1                               0  …                   0.0
2                               1  …                   0.0
3                               1  …                   0.0
4                               0  …                   0.0
..                            …  …                   …
205                             0  …                   0.0
206                             1  …                   0.0
207                             1  …                   0.0
208                             1  …                   0.0
209                             0  …                   0.0

     Days Admitted Admit 4  Total Days Admitted  Total Number of Admissions  \
0                      0.0                    3                           1
1                      0.0                    0                           0
2                      0.0                    0                           0
3                      0.0                    0                           0
4                      0.0                    0                           0
```

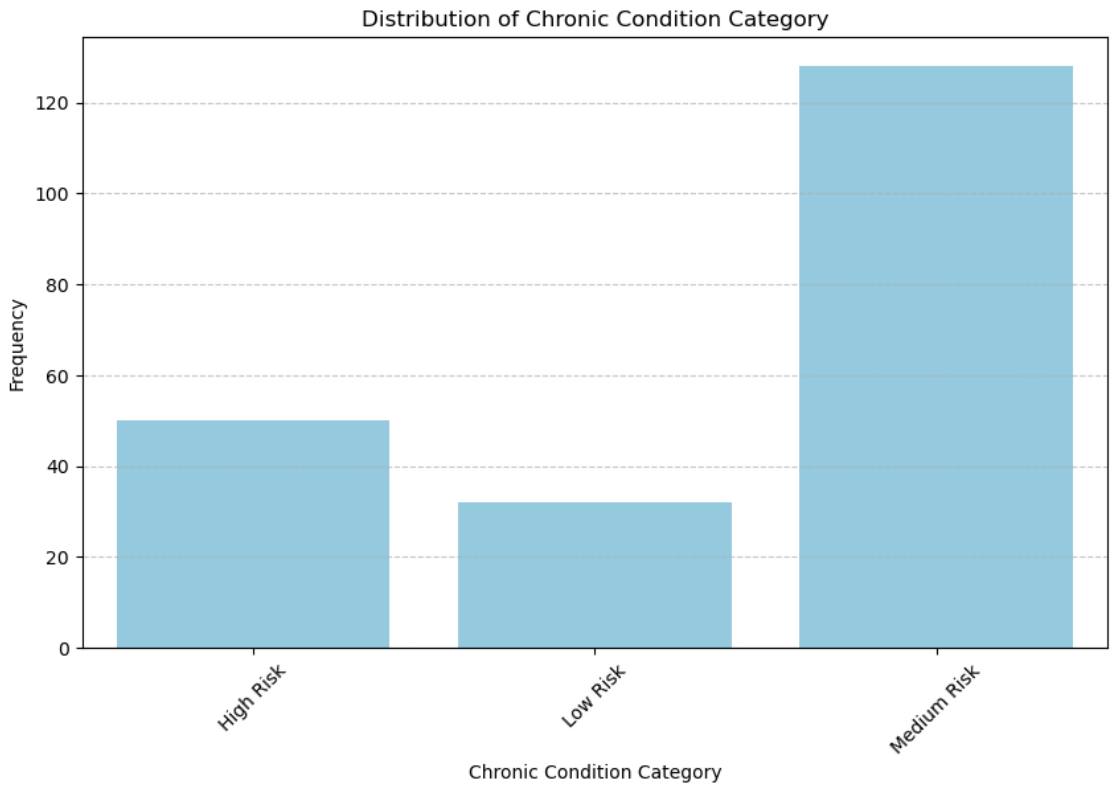
```
..                      …                    …                          …
205                   0.0                    0                          0
206                   0.0                    0                          0
207                   0.0                    8                          1
208                   0.0                    0                          0
209                   0.0                    0                          0

     Change in Weight  Weight change as percent  Age Group  Health Risk Score  \
0                55.8                     20.07          0              49.07
1                -4.8                     -1.70          1              32.30
2                -9.2                     -5.16          2              29.84
3                -5.0                     -2.29          4              63.21
4                 7.5                      3.29          2              32.29
..                 …                        …         …                  …
205               NaN                      0.00          4              45.00
206              -4.1                     -1.94          1              33.06
207              -0.4                     -0.21          4              65.79
208               4.4                      1.91          4              59.41
209             -20.4                     -7.72          3              41.78

     Vaccine History Score  Chronic Condition Category
0                      1.5                 Medium Risk
1                      1.5                 Medium Risk
2                      2.5                 Medium Risk
3                      2.0                   High Risk
4                      0.5                    Low Risk
..                      …                          …
205                    2.0                 Medium Risk
206                    2.5                 Medium Risk
207                    2.0                   High Risk
208                    2.0                   High Risk
209                    2.0                   High Risk

[210 rows x 68 columns]
```

2 Engineered Feature Visualizations

[143]:

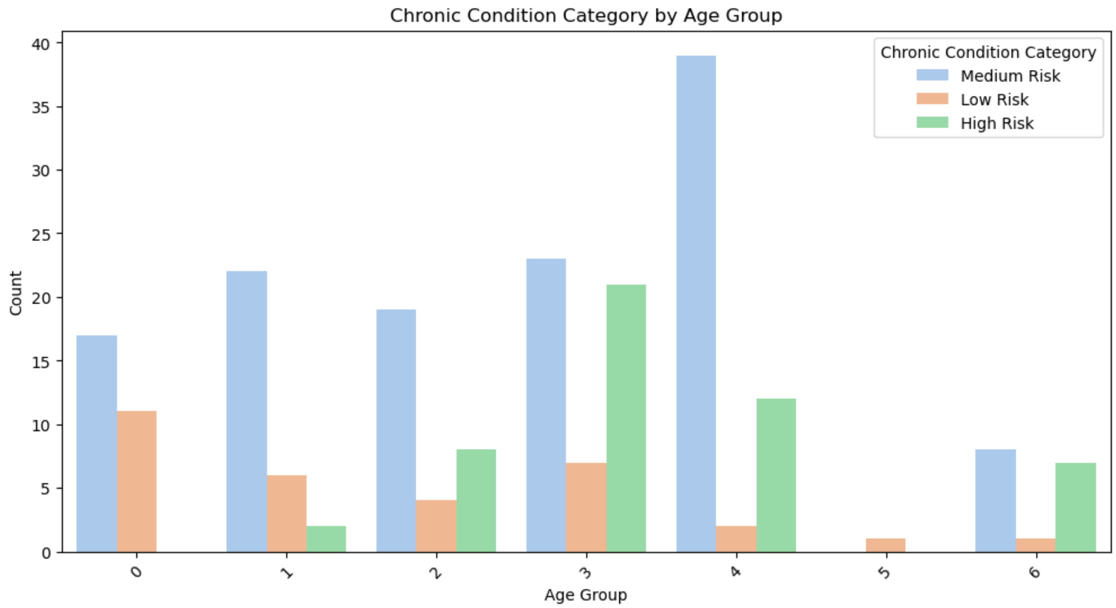
```python
# Aggregate data to count occurrences for each category
data_counts = df_numerical['Chronic Condition Category'].value_counts().
 ↪sort_index()

# Create a bar plot
plt.figure(figsize=(10, 6))
sns.barplot(x=data_counts.index, y=data_counts.values, color='skyblue')

# Customize the plot
```

```
plt.title('Distribution of Chronic Condition Category')
plt.xlabel('Chronic Condition Category')
plt.ylabel('Frequency')
plt.xticks(rotation=45)
plt.grid(axis='y', linestyle='--', alpha=0.7)  # Optional: Add grid lines for
 ↪clarity

plt.show()
```



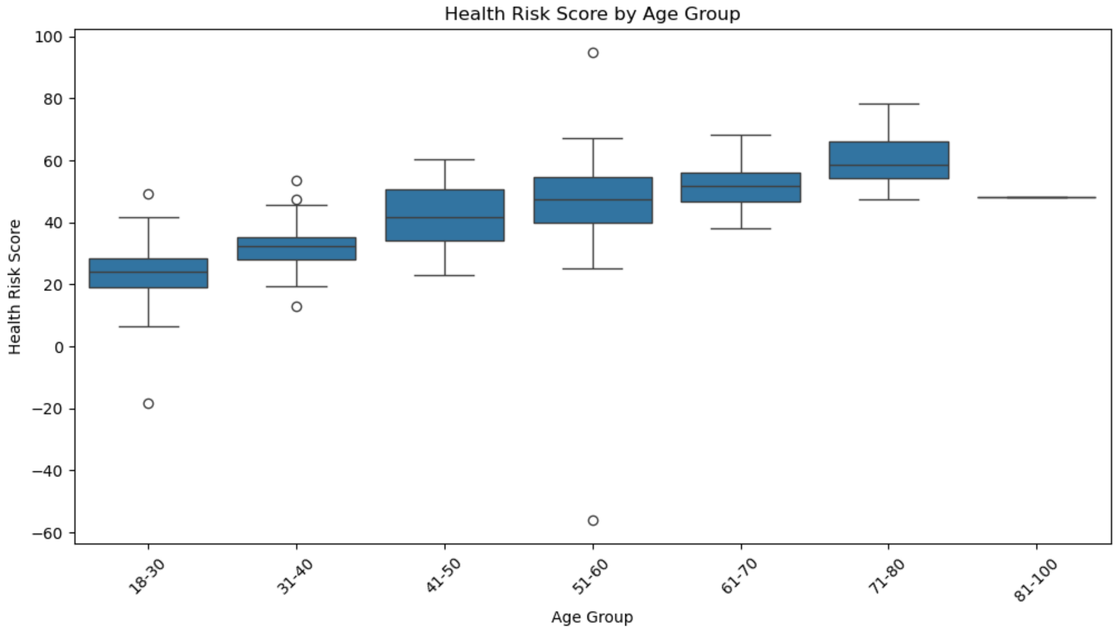
2.0.1 Chronic Condition Categories by Age Group

[150]:
```
# Bar chart of Chronic Condition Category by Age Group
plt.figure(figsize=(12, 6))
sns.countplot(x='Age Group', hue='Chronic Condition Category',
 ↪data=df_numerical, palette='pastel')
plt.title('Chronic Condition Category by Age Group')
plt.xlabel('Age Group')
plt.ylabel('Count')
plt.xticks(rotation=45)  # Rotate labels for readability
plt.legend(title='Chronic Condition Category', loc='upper right')
plt.show()
```



2.0.2 Health Risk Score by Age Group

[153]:
```python
# Boxplot of Health Risk Score by Age Group
plt.figure(figsize=(12, 6))
sns.boxplot(x='Age Group', y='Health Risk Score', data=df)
plt.title('Health Risk Score by Age Group')
plt.xlabel('Age Group')
plt.ylabel('Health Risk Score')
plt.xticks(rotation=45)  # Rotate labels for readability
plt.show()
```

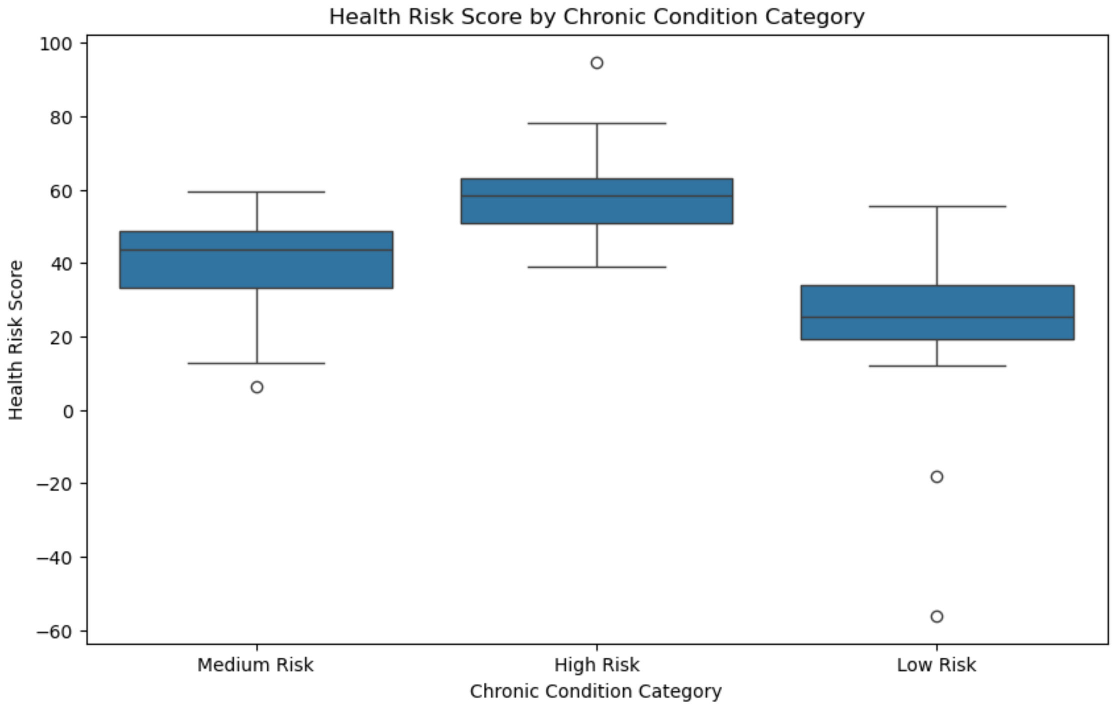


2.0.3 Health Risk Score vs Total Infections

[156]:
```
# Scatter plot for Health Risk Score vs Total Infections
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Health Risk Score', y='Total Infections Since 2022',␣
↪data=df_numerical, hue='Chronic Condition Category', palette='coolwarm')
plt.title('Health Risk Score vs Total Infections')
plt.xlabel('Health Risk Score')
plt.ylabel('Total Infections Since 2022')
plt.legend(title='Chronic Condition Category')
plt.show()
```

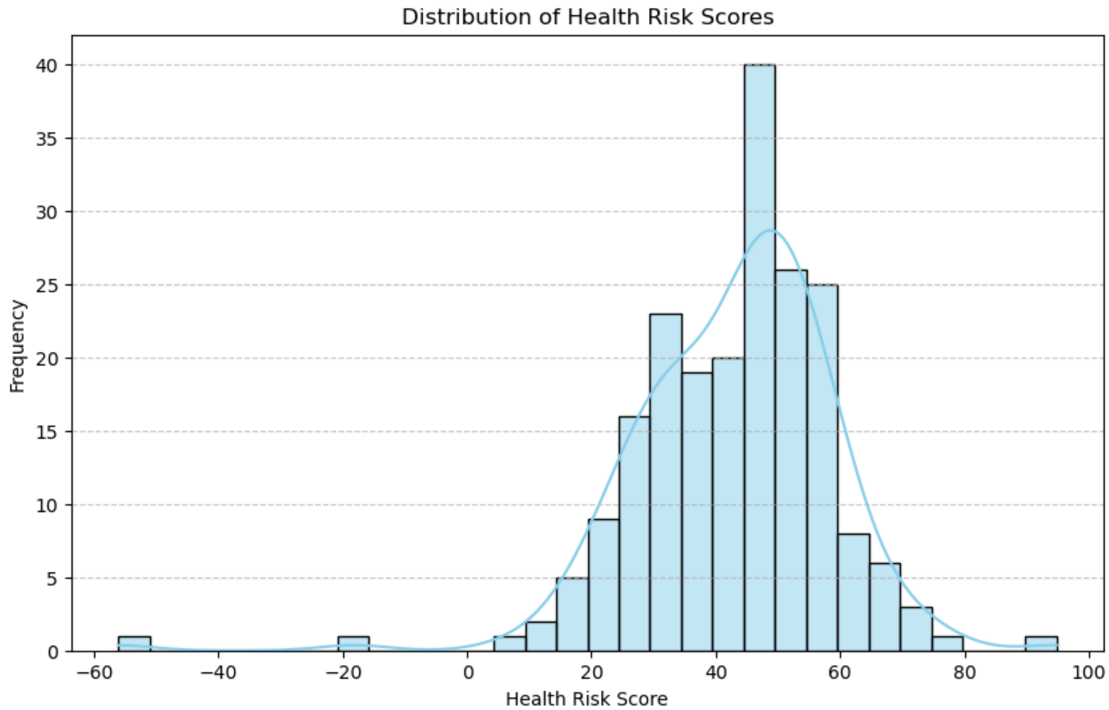
2.0.4 Health Risk Score Distribution by Chronic Condition Category

[159]:
```
# Boxplot of Health Risk Score by Chronic Condition Category
plt.figure(figsize=(10, 6))
sns.boxplot(x='Chronic Condition Category', y='Health Risk Score',␣
↪data=df_numerical)
plt.title('Health Risk Score by Chronic Condition Category')
plt.xlabel('Chronic Condition Category')
plt.ylabel('Health Risk Score')
plt.show()
```



[161]:

```python
# Create a histogram with a KDE overlay for Health Risk Scores
plt.figure(figsize=(10, 6))
sns.histplot(df_numerical['Health Risk Score'], kde=True, bins=30,␣
↪color='skyblue')
plt.title('Distribution of Health Risk Scores')
plt.xlabel('Health Risk Score')
plt.ylabel('Frequency')
plt.grid(axis='y', linestyle='--', alpha=0.7)  # Optional: Add grid lines for␣
↪clarity
plt.show()
```



[163]:
```
df.head()
```

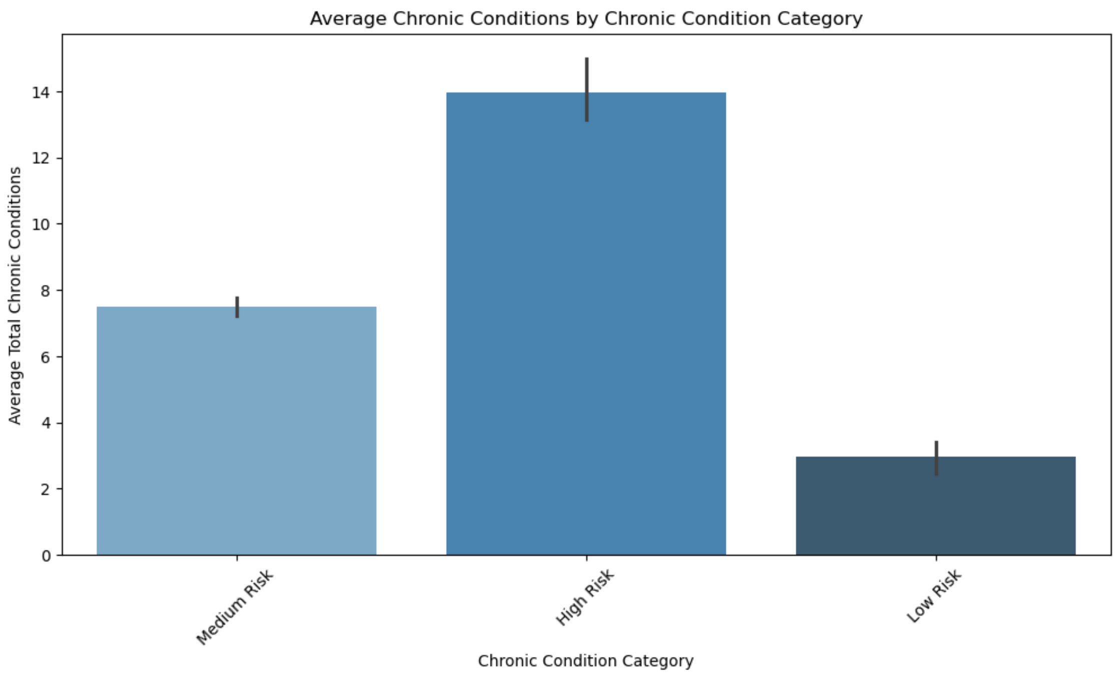
[163]:
```
   Resident ID Gender at Birth   Room  Floor        DOB  AGE  \
0            1          Female  325-2      3 1994-01-01   30
1            2          Female  229-1      2 1992-03-31   32
2            3            Male  426-2      4 1974-11-12   50
3            4          Female  301-1      3 1963-09-11   61
4          227            Male  306-1      3 1978-08-14   46

   Total Antibiotics COVID Pandemic Vaccine #1 Date  \
0                  0                     2021-12-15
1                  0                     2021-07-01
2                  0                     2021-01-03
3                  0                     2021-01-06
4                  0                            NaT

  COVID Pandemic Vaccine #2 Date 21 - 22 COVID Vaccination Status  …  \
0                     2022-01-24                       Up To Date  …
1                     2021-07-01                       Up To Date  …
2                     2021-01-31                       Up To Date  …
3                     2021-02-03                       Up To Date  …
4                            NaT                         Declined  …

  Days Admitted Admit3 Admit Date 4 Diagnosis Admit  4 Days Admitted Admit 4  \
```

```
0                   NaN          NaT                 NaN                    NaN
1                   NaN          NaT                 NaN                    NaN
2                   NaN          NaT                 NaN                    NaN
3                   NaN          NaT                 NaN                    NaN
4                   NaN          NaT                 NaN                    NaN

   Total Days Admitted Total Number of Admissions Change in Weight  \
0                    3                          1             55.8
1                    0                          0             -4.8
2                    0                          0             -9.2
3                    0                          0               -5
4                    0                          0              7.5

  Weight change as percent Age Group Health Risk Score
0                    20.07     18-30             49.07
1                    -1.70     31-40             32.30
2                    -5.16     41-50             29.84
3                    -2.29     61-70             63.21
4                     3.29     41-50             32.29

[5 rows x 85 columns]
```

2.0.5 Total Chronic Conditions by Chronic Illness Category

[166]:

```python
# Total Chronic conditions by Total Chronic Conditions

# Create the bar chart
plt.figure(figsize=(12, 6))

sns.barplot(
    x='Chronic Condition Category',
    y='Total Chronic Conditions',
    hue='Chronic Condition Category',  # Explicitly use 'Age Group' for␣
 ↪grouping (if applicable)
    data=df_numerical,
    palette="Blues_d",
    dodge=False  # Use dodge=False for single-grouped bars
)

plt.title('Average Chronic Conditions by Chronic Condition Category')
plt.xlabel('Chronic Condition Category')
plt.ylabel('Average Total Chronic Conditions')
plt.xticks(rotation=45)
plt.show()
```



[168]:
```
df_numerical['Vaccine History Score']
```

[168]:
```
0      1.5
1      1.5
2      2.5
3      2.0
4      0.5
      …
205    2.0
206    2.5
207    2.0
208    2.0
209    2.0
Name: Vaccine History Score, Length: 210, dtype: float64
```

[170]:
```
# Add Engineered Featured to Scaled Df

df_scaled[['Health Risk Score', 'Vaccine History Score', 'Chronic Condition␣
↪Category']] = df_numerical[['Health Risk Score', 'Vaccine History Score',␣
↪'Chronic Condition Category']]
```

[172]:
```
df_scaled.head()
```

[172]:
```
   Resident ID  Gender at Birth  Floor       AGE  Total Antibiotics  \
0            1                0      3 -1.408157          -0.433781
```

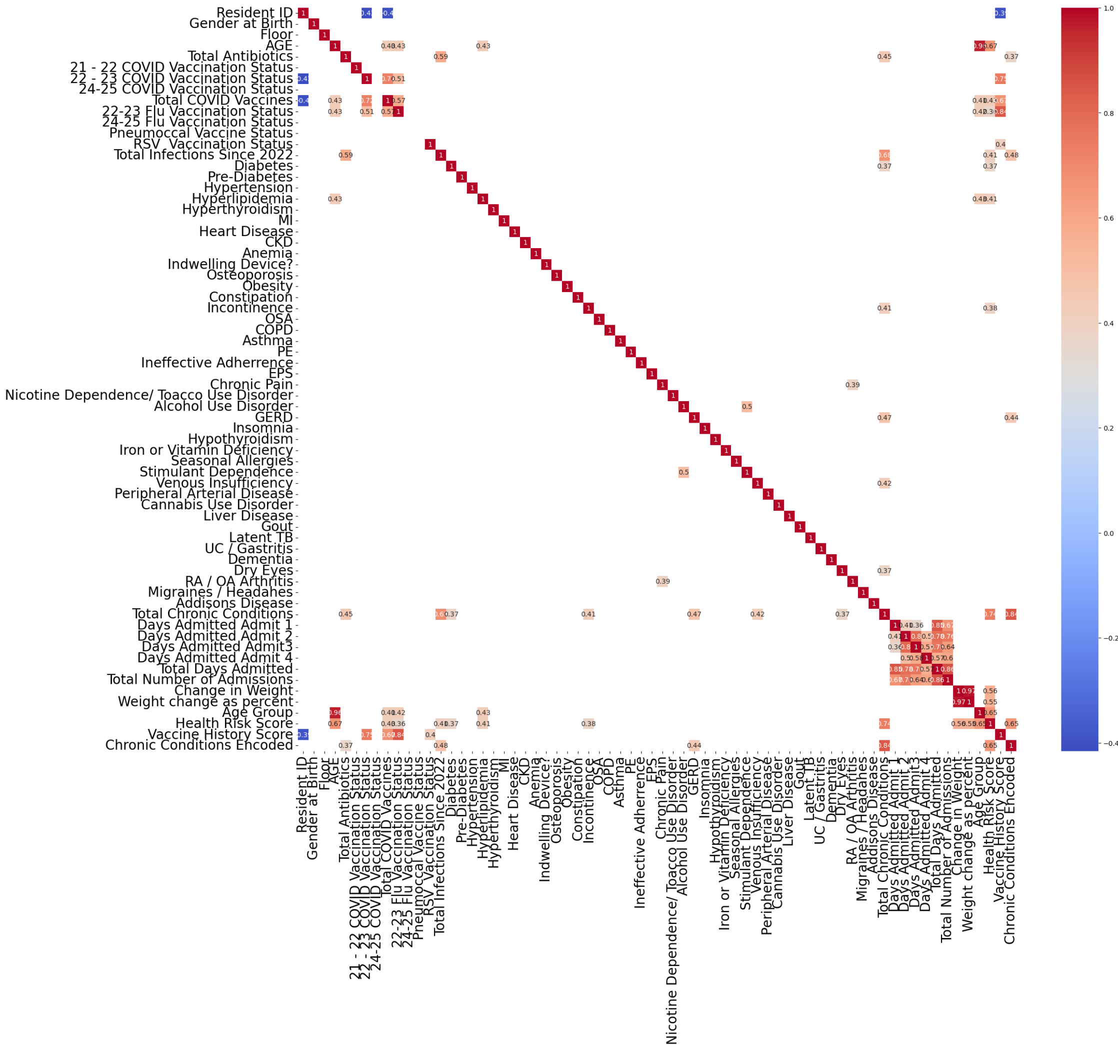
```
1           2                0      2 -1.275936          -0.433781
2           3                1      4 -0.085944          -0.433781
3           4                0      3  0.641273          -0.433781
4         227                1      3 -0.350387          -0.433781

   21 - 22 COVID Vaccination Status  22 - 23 COVID Vaccination Status  \
0                               0.0                                 1
1                               0.0                                 1
2                               0.0                                 1
3                               0.0                                 1
4                               0.0                                 0

   24-25 COVID Vaccination Status  Total COVID Vaccines  \
0                               0             -0.049371
1                               0              0.814629
2                               0              0.814629
3                               0              0.814629
4                               0             -1.777371

   22-23 Flu Vaccination Status  …  Days Admitted Admit3  \
0                             0  …             -0.155292
1                             0  …             -0.155292
2                             1  …             -0.155292
3                             1  …             -0.155292
4                             0  …             -0.155292

   Days Admitted Admit 4  Total Days Admitted  Total Number of Admissions  \
0              -0.131533             0.450674                    0.817841
1              -0.131533            -0.365202                   -0.435784
2              -0.131533            -0.365202                   -0.435784
3              -0.131533            -0.365202                   -0.435784
4              -0.131533            -0.365202                   -0.435784

   Change in Weight  Weight change as percent  Age Group  Health Risk Score  \
0          3.446519                  2.484106          0              49.07
1         -0.256260                 -0.176451          1              32.30
2         -0.525109                 -0.599305          2              29.84
3         -0.268481                 -0.248556          4              63.21
4          0.495294                  0.433387          2              32.29

   Vaccine History Score  Chronic Condition Category
0                    1.5                 Medium Risk
1                    1.5                 Medium Risk
2                    2.5                 Medium Risk
3                    2.0                   High Risk
4                    0.5                    Low Risk
```

[5 rows x 68 columns]

2.0.6 Correlation Heatmap

[175]:

```
# Correlation Heatmap

heatmap_df = df_numerical.drop(columns=['Chronic Condition Category'])

corr_matrix = heatmap_df.corr()

# Create a mask for correlations less than 0.5
mask = np.abs(corr_matrix) < 0.35

plt.figure(figsize=(24, 20))
sns.heatmap(corr_matrix, annot=True, cmap="coolwarm", mask=mask)
plt.xticks(rotation=90, fontsize = 20)
plt.yticks(rotation=0, fontsize = 20)
plt.show()
```



2.1 Feature Selection

- Final DataFrame will be df_final

[ ]:

```

```