cwy_ 1_part1

March 27, 2023

[ ]:

```
# Alzheimer Disease and Healthy Aging Data In US
!wget -O data/data.zip "https://storage.googleapis.com/kaggle-data-sets/2996762/
 ↪5157319/compressed/
 ↪Alzheimer%20Disease%20and%20Healthy%20Aging%20Data%20In%20US.csv.zip?
 ↪X-Goog-Algorithm=GOOG4-RSA-SHA256&X-Goog-Credential=gcp-kaggle-com%40kaggle-161607.
 ↪iam.gserviceaccount.
 ↪com%2F20230327%2Fauto%2Fstorage%2Fgoog4_request&X-Goog-Date=20230327T090318Z&X-Goog-Expires
!unzip -o data/data.zip -d data/
```

```
--2023-03-27 21:54:32--  https://storage.googleapis.com/kaggle-data-sets/2996762
/5157319/compressed/Alzheimer%20Disease%20and%20Healthy%20Aging%20Data%20In%20US
.csv.zip?X-Goog-Algorithm=GOOG4-RSA-SHA256&X-Goog-Credential=gcp-kaggle-com%40ka
ggle-161607.iam.gserviceaccount.com%2F20230327%2Fauto%2Fstorage%2Fgoog4_request&
X-Goog-Date=20230327T090318Z&X-Goog-Expires=259200&X-Goog-
SignedHeaders=host&X-Goog-Signature=cbdfd3404c8b0b63dfa1c8eb7295c847335720b2bfed
1403fbe8222101a03938b5c984bd26672ed6be52a597b77d347bc2c06c93a9c7f5e45fc0941fa868
b1111f1c7a5c08b6d646db7849289b1fbf14539e9b311a66ff1abae1d091fa66cf777fef79854710
92b773c7602db4c758abf895088998f0821cca2bf01f565c0323b354a08a82a166837cbb5de07ec4
7bb96b43da5152788ca4fe1137cd4166ca170572b8ca8e05b86f44c6e8c26afedf7c0d0b262bf298
c4eb643c8239e1a4757bfc37cb02cd15057291f5ec9dcda3a38752a395fdb7ce30904cffacf8a572
b6306834278c4383ec6fd03d65360ab006af754826c1a0e3e0b52259452018df9791
Resolving storage.googleapis.com (storage.googleapis.com)… 172.217.163.48,
142.251.43.16, 142.251.42.240, …
Connecting to storage.googleapis.com
(storage.googleapis.com)|172.217.163.48|:443… connected.
HTTP request sent, awaiting response… 200 OK
Length: 5460031 (5.2M) [application/zip]
Saving to: ‘data/data.zip’

data/data.zip       100%[===================>]   5.21M  10.5MB/s    in 0.5s

2023-03-27 21:54:34 (10.5 MB/s) - ‘data/data.zip’ saved [5460031/5460031]

Archive:  data/data.zip
  inflating: data/Alzheimer Disease and Healthy Aging Data In US.csv
```

[ ]:

```python
import pandas as pd
import matplotlib.pyplot as plt

#
df = pd.read_csv("data/Alzheimer Disease and Healthy Aging Data In US.csv")

print(df.head(5))
#
print("   ")
print(df.describe())
```

```
/tmp/ipykernel_2201461/2559574171.py:5: DtypeWarning: Columns (13,14) have mixed
types. Specify dtype option on import or set low_memory=False.
  df = pd.read_csv("data/Alzheimer Disease and Healthy Aging Data In US.csv")

   YearStart  YearEnd LocationAbbr LocationDesc Datasource           Class  \
0       2020     2020           HI       Hawaii      BRFSS  Overall Health
1       2017     2017           ID        Idaho      BRFSS   Mental Health
2       2017     2017           ID        Idaho      BRFSS  Overall Health
3       2018     2018           ID        Idaho      BRFSS  Overall Health
4       2020     2020           IN      Indiana      BRFSS   Mental Health

                                             Topic  \
0                     Arthritis among older adults
1                 Lifetime diagnosis of depression
2                     Arthritis among older adults
3  Physically unhealthy days (mean number of days)
4                 Lifetime diagnosis of depression

                                            Question Data_Value_Unit  \
0  Percentage of older adults ever told they have…               %
1  Percentage of older adults with a lifetime dia…               %
2  Percentage of older adults ever told they have…               %
3  Physically unhealthy days (mean number of days…          Number
4  Percentage of older adults with a lifetime dia…               %

  DataValueTypeID  … Stratification2                      Geolocation  \
0           PRCTG  …             NaN  POINT (-157.8577494 21.30485044)
1           PRCTG  …             NaN    POINT (-114.36373 43.68263001)
2           PRCTG  …             NaN    POINT (-114.36373 43.68263001)
3            MEAN  …             NaN    POINT (-114.36373 43.68263001)
4           PRCTG  …            Male  POINT (-86.14996019 39.76691045)

   ClassID TopicID QuestionID  LocationID StratificationCategoryID1  \
0      C01   TOC11        Q43          15                       AGE
1      C05   TMC03        Q27          16                       AGE
2      C01   TOC11        Q43          16                       AGE
3      C01   TOC01        Q08          16                       AGE
```

```
4      C05    TMC03         Q27          18                        AGE

  StratificationID1 StratificationCategoryID2 StratificationID2
0              5064                   OVERALL           OVERALL
1              5064                   OVERALL           OVERALL
2              5064                   OVERALL           OVERALL
3              5064                   OVERALL           OVERALL
4       AGE_OVERALL                    GENDER              MALE

[5 rows x 29 columns]

           YearStart        YearEnd     Data_Value  Data_Value_Alt  \
count  214462.000000  214462.000000  144629.000000   144629.000000
mean     2017.378477    2017.634000      37.341956       37.341956
std         1.779822       1.778926      25.183017       25.183017
min      2015.000000    2015.000000       0.000000        0.000000
25%      2016.000000    2016.000000      15.300000       15.300000
50%      2017.000000    2018.000000      32.500000       32.500000
75%      2019.000000    2019.000000      56.800000       56.800000
max      2020.000000    2020.000000     100.000000      100.000000

       Sample_Size     LocationID
count          0.0  214462.000000
mean           NaN     800.987821
std            NaN    2512.934094
min            NaN       1.000000
25%            NaN      18.000000
50%            NaN      33.000000
75%            NaN      49.000000
max            NaN    9004.000000
```

[ ]:

```python
#
for column in df.columns:
    if df[column].dtype == 'object':
        print(f"\n   {column}   ")
        print(df[column].value_counts())
```

```
   LocationAbbr
US       4644
WEST     4638
NRE      4614
MDW      4611
OR       4565
NY       4557
SOU      4542
UT       4222
OH       3955
```

GA	3951
MD	3919
HI	3907
TN	3879
MI	3796
VA	3758
FL	3753
ME	3733
TX	3699
NV	3696
DC	3684
WV	3682
MS	3677
PA	3648
NM	3635
AL	3633
KY	3623
AK	3611
SC	3592
NJ	3589
AZ	3582
MO	3573
IL	3571
IN	3570
WI	3569
LA	3563
MN	3555
NE	3546
CT	3543
RI	3534
OK	3526
SD	3526
ND	3514
KS	3510
ID	3507
IA	3501
AR	3498
WY	3494
CA	3447
CO	3390
NC	3349
WA	3348
MT	3348
DE	3346
NH	3284
VT	3278
MA	3174
PR	2797

```
GU      2703
VI       503
Name: LocationAbbr, dtype: int64

   LocationDesc
United States, DC & Territories    4644
West                               4638
Northeast                          4614
Midwest                            4611
Oregon                             4565
New York                           4557
South                              4542
Utah                               4222
Ohio                               3955
Georgia                            3951
Maryland                           3919
Hawaii                             3907
Tennessee                          3879
Michigan                           3796
Virginia                           3758
Florida                            3753
Maine                              3733
Texas                              3699
Nevada                             3696
District of Columbia               3684
West Virginia                      3682
Mississippi                        3677
Pennsylvania                       3648
New Mexico                         3635
Alabama                            3633
Kentucky                           3623
Alaska                             3611
South Carolina                     3592
New Jersey                         3589
Arizona                            3582
Missouri                           3573
Illinois                           3571
Indiana                            3570
Wisconsin                          3569
Louisiana                          3563
Minnesota                          3555
Nebraska                           3546
Connecticut                        3543
Rhode Island                       3534
Oklahoma                           3526
South Dakota                       3526
North Dakota                       3514
Kansas                             3510
```

```
Idaho                                  3507
Iowa                                   3501
Arkansas                               3498
Wyoming                                3494
California                             3447
Colorado                               3390
North Carolina                         3349
Washington                             3348
Montana                                3348
Delaware                               3346
New Hampshire                          3284
Vermont                                3278
Massachusetts                          3174
Puerto Rico                            2797
Guam                                   2703
Virgin Islands                          503
Name: LocationDesc, dtype: int64

   Datasource
BRFSS    214462
Name: Datasource, dtype: int64

   Class
Overall Health                            71694
Screenings and Vaccines                   46867
Nutrition/Physical Activity/Obesity       24851
Cognitive Decline                         19180
Caregiving                                18671
Mental Health                             16600
Smoking and Alcohol Use                   16599
Name: Class, dtype: int64

   Topic
Obesity
8300
Influenza vaccine within past year
8300
Physically unhealthy days (mean number of days)
8300
Frequent mental distress
8300
Current smoking
8300
Lifetime diagnosis of depression
8300
No leisure-time physical activity within past month
8300
Self-rated health (fair to poor health)
```

```
8299
Self-rated health (good to excellent health)
8299
Binge drinking within past 30 days
8299
Ever had pneumococcal vaccine
8268
Recent activity limitations in past month
8233
Disability status, including sensory or mobility limitations
6917
Arthritis among older adults
5511
Fair or poor health among older adults with arthritis
5447
Subjective cognitive decline or memory loss among older adults
5088
Diabetes screening within past 3 years
4808
Talked with health care professional about subjective cognitive decline or
memory loss                    4700
Need assistance with day-to-day activities because of subjective cognitive
decline or memory loss         4696
Functional difficulties associated with subjective cognitive decline or memory
loss among older adults    4696
Fall with injury within last year
4173
Colorectal cancer screening
4173
Oral health:  tooth retention
4172
Prevalence of sufficient sleep
4171
Eating 3 or more vegetables daily
4127
High blood pressure ever
4127
Cholesterol checked in past 5 years
4127
Eating 2 or more fruits daily
4124
Taking medication for high blood pressure
4108
Severe joint pain among older adults with arthritis
4064
Provide care for a friend or family member in past month
3848
Expect to provide care for someone in the next two years
```

```
3797
Provide care for someone with cognitive impairment within the past month
3682
Duration of caregiving among older adults
3681
Intensity of caregiving among older adults
3663
Up-to-date with recommended vaccines and screenings - Women
3280
Up-to-date with recommended vaccines and screenings - Men
3271
Mammogram within past 2 years
3271
Pap test within past 3 years
3242
Name: Topic, dtype: int64

   Question
Percentage of older adults who are currently obese, with a body mass index (BMI)
of 30 or more
8300
Percentage of older adults who reported influenza vaccine within the past year
8300
Physically unhealthy days (mean number of days in past month)
8300
Percentage of older adults who are experiencing frequent mental distress
8300
Percentage of older adults who have smoked at least 100 cigarettes in their
entire life and still smoke every day or some days
8300
Percentage of older adults with a lifetime diagnosis of depression
8300
Percentage of older adults who have not had any leisure time physical activity
in the past month
8300
Percentage of older adults who self-reported that their health is "fair" or
"poor"
8299
Percentage of older adults who self-reported that their health is "good", "very
good", or "excellent"
8299
Percentage of older adults who reported binge drinking within the past 30 days
8299
Percentage of at risk adults (have diabetes, asthma, cardiovascular disease or
currently smoke) who ever had a pneumococcal vaccine
8268
Mean number of days with activity limitations in the past month
8233
```

```
Percentage of older adults who report having a disability (includes limitations
related to sensory or mobility impairments or a physical, mental, or emotional
condition)    6917
Percentage of older adults ever told they have arthritis
5511
Fair or poor health among older adults with doctor-diagnosed arthritis
5447
Percentage of older adults who reported subjective cognitive decline or memory
loss that is happening more often or is getting worse in the preceding 12 months
5088
Percentage of older adults without diabetes who reported a blood sugar or
diabetes test within 3 years
4808
Percentage of older adults with subjective cognitive decline or memory loss who
reported talking with a health care professional about it
4700
Percentage of older adults who reported that as a result of subjective cognitive
decline or memory loss that they need assistance with day-to-day activities
4696
Percentage of older adults who reported subjective cognitive decline or memory
loss that interferes with their ability to engage in social activities or
household chores    4696
Percentage of older adults who have fallen and sustained an injury within last
year
4173
Percentage of older adults who had either a home blood stool test within the
past year or a sigmoidoscopy or colonoscopy within the past 10 years
4173
Percentage of older adults who report having lost 5 or fewer teeth due to decay
or gum disease
4172
Percentage of older adults getting sufficient sleep (>6 hours)
4171
Percentage of older adults who are eating 3 or more vegetables daily
4127
Percentage of older adults who have ever been told by a health professional that
they have high blood pressure
4127
Percentage of older adults who had a cholesterol screening within the past 5
years
4127
Percentage of older adults who are eating 2 or more fruits daily
4124
Percentage of older adults who have been told they have high blood pressure who
report currently taking medication for their high blood pressure
4108
Severe joint pain due to arthritis among older adults with doctor-diagnosed
arthritis
```

```
4064
Percentage of older adults who provided care for a friend or family member
within the past month
3848
Percentage of older adults currently not providing care who expect to provide
care for someone with health problems in the next two years
3797
Percentage of older adults who provided care for someone with dementia or other
cognitive impairment within the past month
3682
Percentage of older adults who provided care to a friend or family member for
six months or more
3681
Average of 20 or more hours of care per week provided to a friend or family
member
3663
Percentage of older adult women who are up to date with select clinical
preventive services
3280
Percentage of older adult men who are up to date with select clinical preventive
services
3271
Percentage of older adult women who have received a mammogram within the past 2
years
3271
Percentage of older adult women with an intact cervix who had a Pap test within
the past 3 years
3242
Name: Question, dtype: int64

   Data_Value_Unit
%         197929
Number     16533
Name: Data_Value_Unit, dtype: int64

   DataValueTypeID
PRCTG    197929
MEAN      16533
Name: DataValueTypeID, dtype: int64

   Data_Value_Type
Percentage    197929
Mean           16533
Name: Data_Value_Type, dtype: int64

   Low_Confidence_Limit
5.4     350
5.1     318
```

```
4.8     314
5.3     313
5       305
       …
96.8      1
98.4      1
0.9       1
97.8      1
99.4      1
Name: Low_Confidence_Limit, Length: 1967, dtype: int64

   High_Confidence_Limit
6.5    216
5.8    193
6.8    192
6.7    189
7.5    186
      …
1.7      1
2.8      1
2.8      1
1.5      1
1.6      1
Name: High_Confidence_Limit, Length: 1966, dtype: int64

   StratificationCategory1
Age Group    214462
Name: StratificationCategory1, dtype: int64

   Stratification1
Overall              71919
50-64 years          71528
65 years or older    71015
Name: Stratification1, dtype: int64

   StratificationCategory2
Race/Ethnicity    134959
Gender             51834
Name: StratificationCategory2, dtype: int64

   Stratification2
White, non-Hispanic         27633
Hispanic                    27525
Black, non-Hispanic         26968
Native Am/Alaskan Native    26571
Asian/Pacific Islander      26262
Female                      26091
Male                        25743
```

```
Name: Stratification2, dtype: int64

   Geolocation
POINT (-120.1550313 44.56744942)    4565
POINT (-75.54397043 42.82700103)    4557
POINT (-111.5871306 39.36070017)    4222
POINT (-82.40426006 40.06021014)    3955
POINT (-83.62758035 32.83968109)    3951
POINT (-76.60926011 39.29058096)    3919
POINT (-157.8577494 21.30485044)    3907
POINT (-85.77449091 35.68094058)    3879
POINT (-84.71439027 44.66131954)    3796
POINT (-78.45789046 37.54268067)    3758
POINT (-81.92896054 28.93204038)    3753
POINT (-68.98503134 45.25422889)    3733
POINT (-99.42677021 31.82724041)    3699
POINT (-117.0718406 39.49324039)    3696
POINT (-77.036871 38.907192)        3684
POINT (-80.71264013 38.6655102)     3682
POINT (-89.53803082 32.7455101)     3677
POINT (-77.86070029 40.79373015)    3648
POINT (-106.240581 34.52088095)     3635
POINT (-86.63186076 32.84057112)    3633
POINT (-84.77497105 37.64597027)    3623
POINT (-147.722059 64.84507996)     3611
POINT (-81.04537121 33.9988213)     3592
POINT (-74.27369129 40.13057005)    3589
POINT (-111.7638113 34.86597028)    3582
POINT (-92.56630005 38.63579078)    3573
POINT (-88.99771018 40.48501028)    3571
POINT (-86.14996019 39.76691045)    3570
POINT (-89.81637074 44.39319117)    3569
POINT (-92.44568007 31.31266064)    3563
POINT (-94.7942005 46.35564874)     3555
POINT (-99.36572062 41.64104099)    3546
POINT (-72.64984095 41.56266102)    3543
POINT (-71.52247031 41.70828019)    3534
POINT (-97.52107021 35.47203136)    3526
POINT (-100.3735306 44.35313005)    3526
POINT (-100.118421 47.47531978)     3514
POINT (-98.20078123 38.3477403)     3510
POINT (-114.36373 43.68263001)      3507
POINT (-93.81649056 42.46940091)    3501
POINT (-92.27449074 34.74865012)    3498
POINT (-108.1098304 43.23554134)    3494
POINT (-120.9999995 37.63864012)    3447
POINT (-106.1336109 38.84384076)    3390
POINT (-79.15925046 35.46622098)    3349
POINT (-109.4244206 47.06652897)    3348
POINT (-120.4700108 47.52227863)    3348
POINT (-75.57774117 39.00883067)    3346
POINT (-71.50036092 43.65595011)    3284
POINT (-72.51764079 43.62538124)    3278
POINT (-72.08269067 42.27687047)    3174
POINT (-66.590149 18.220833)        2797
POINT (144.793731 13.444304)        2703
POINT (-64.896335 18.335765)         503
Name: Geolocation, dtype: int64

   ClassID
C01    71694
C03    46867
C02    24851
C06    19180
C07    18671
C05    16600
C04    16599
Name: ClassID, dtype: int64

   TopicID
TNC04    8300
TSC08    8300
TOC01    8300
TMC01    8300
TAC01    8300
TMC03    8300
TNC03    8300
TOC07    8299
TOC08    8299
TAC03    8299
TSC09    8268
TOC03    8233
TOC10    6917
TOC11    5511
TOC13    5447
TCC01    5088
TSC04    4808
TCC04    4700
TCC03    4696
TCC02    4696
TOC06    4173
TSC02    4173
TOC05    4172
TOC09    4171
TNC02    4127
TSC07    4127
```

```
TSC06    4127
TNC01    4124
TOC04    4108
TOC12    4064
TGC01    3848
TGC02    3797
TGC05    3682
TGC03    3681
TGC04    3663
TSC11    3280
TSC10    3271
TSC01    3271
TSC03    3242
Name: TopicID, dtype: int64

   QuestionID
Q13    8300
Q18    8300
Q08    8300
Q03    8300
Q17    8300
Q27    8300
Q16    8300
Q32    8299
Q33    8299
Q21    8299
Q09    8268
Q35    8233
Q46    6917
Q43    5511
Q45    5447
Q30    5088
Q19    4808
Q42    4700
Q41    4696
Q31    4696
Q05    4173
Q15    4173
Q07    4172
Q34    4171
Q02    4127
Q22    4127
Q14    4127
Q01    4124
Q04    4108
Q44    4064
Q36    3848
Q37    3797
```

```
Q40    3682
Q38    3681
Q39    3663
Q11    3280
Q10    3271
Q12    3271
Q20    3242
Name: QuestionID, dtype: int64

   StratificationCategoryID1
AGE    214462
Name: StratificationCategoryID1, dtype: int64

   StratificationID1
AGE_OVERALL    71919
5064           71528
65PLUS         71015
Name: StratificationID1, dtype: int64

   StratificationCategoryID2
RACE       134959
GENDER      51834
OVERALL     27669
Name: StratificationCategoryID2, dtype: int64

   StratificationID2
OVERALL    27669
WHT        27633
HIS        27525
BLK        26968
NAA        26571
ASN        26262
FEMALE     26091
MALE       25743
Name: StratificationID2, dtype: int64
```

[ ]:

```
#   5
num_attributes = df.select_dtypes(include=['number'])
for column in num_attributes.columns:
    print(f"\n   {column}  5   ")
    print(df[column].describe())
    print(f"  {column}     {df[column].isna().sum()}")
```

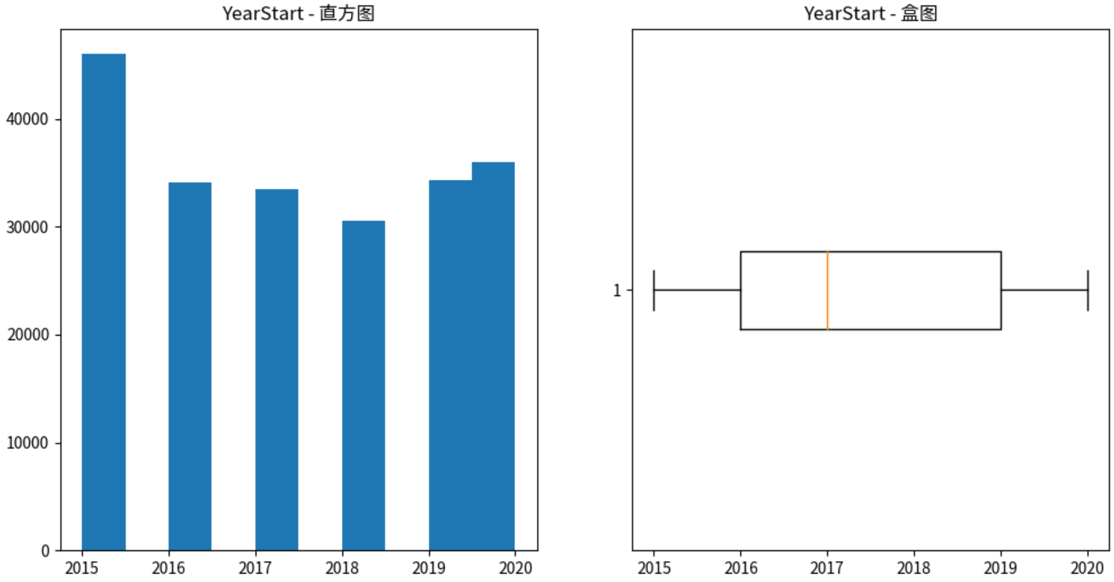
```
   YearStart  5
count    214462.000000
mean       2017.378477
std           1.779822
```

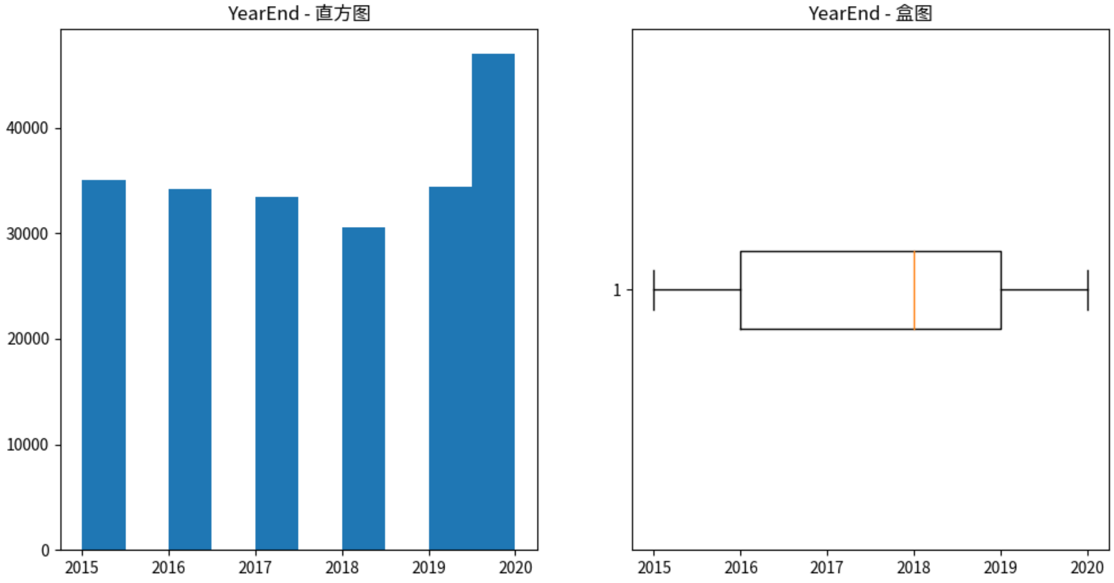
```
min        2015.000000
25%        2016.000000
50%        2017.000000
75%        2019.000000
max        2020.000000
Name: YearStart, dtype: float64
  YearStart     0

   YearEnd  5
count    214462.000000
mean       2017.634000
std           1.778926
min        2015.000000
25%        2016.000000
50%        2018.000000
75%        2019.000000
max        2020.000000
Name: YearEnd, dtype: float64
  YearEnd     0

   Data_Value  5
count    144629.000000
mean         37.341956
std          25.183017
min           0.000000
25%          15.300000
50%          32.500000
75%          56.800000
max         100.000000
Name: Data_Value, dtype: float64
  Data_Value     69833

   Data_Value_Alt  5
count    144629.000000
mean         37.341956
std          25.183017
min           0.000000
25%          15.300000
50%          32.500000
75%          56.800000
max         100.000000
Name: Data_Value_Alt, dtype: float64
  Data_Value_Alt     69833

   Sample_Size  5
count    0.0
mean     NaN
std      NaN
```

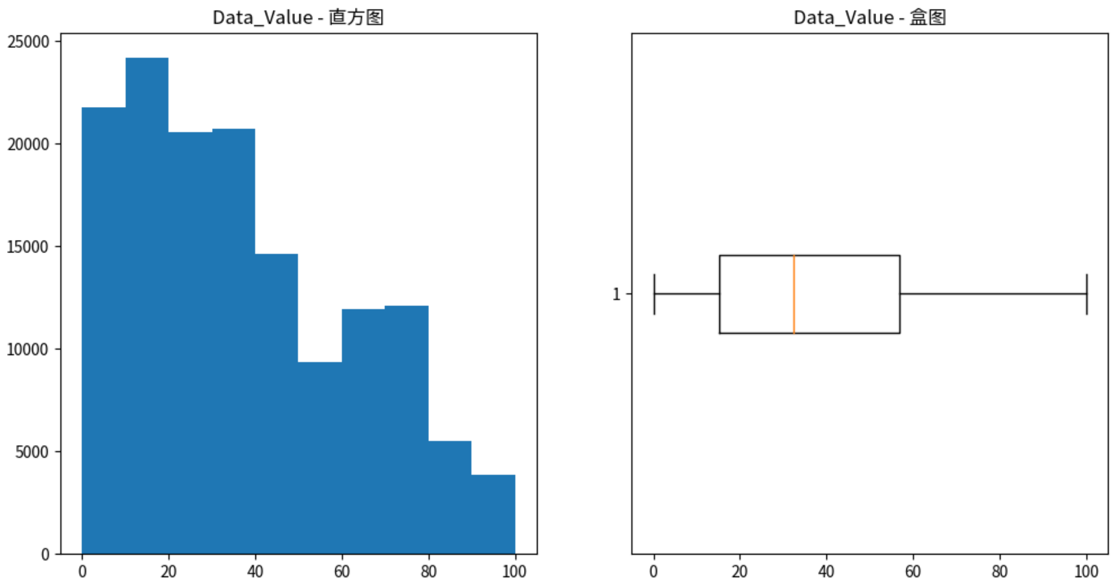
```
min      NaN
25%      NaN
50%      NaN
75%      NaN
max      NaN
Name: Sample_Size, dtype: float64
  Sample_Size     214462

   LocationID  5
count    214462.000000
mean        800.987821
std        2512.934094
min           1.000000
25%          18.000000
50%          33.000000
75%          49.000000
max        9004.000000
Name: LocationID, dtype: float64
  LocationID     0
```

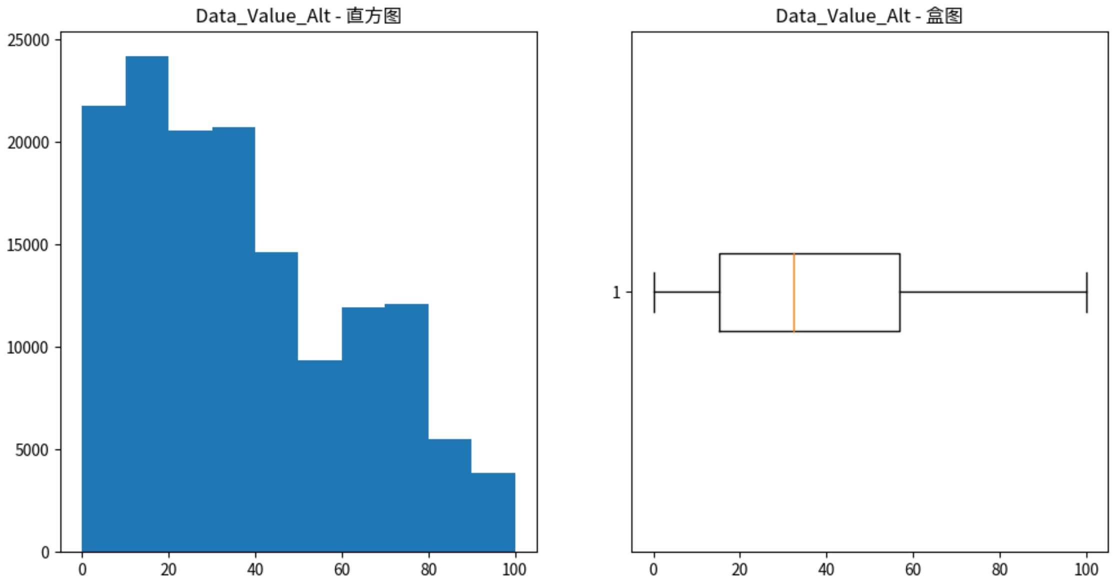
[ ]:

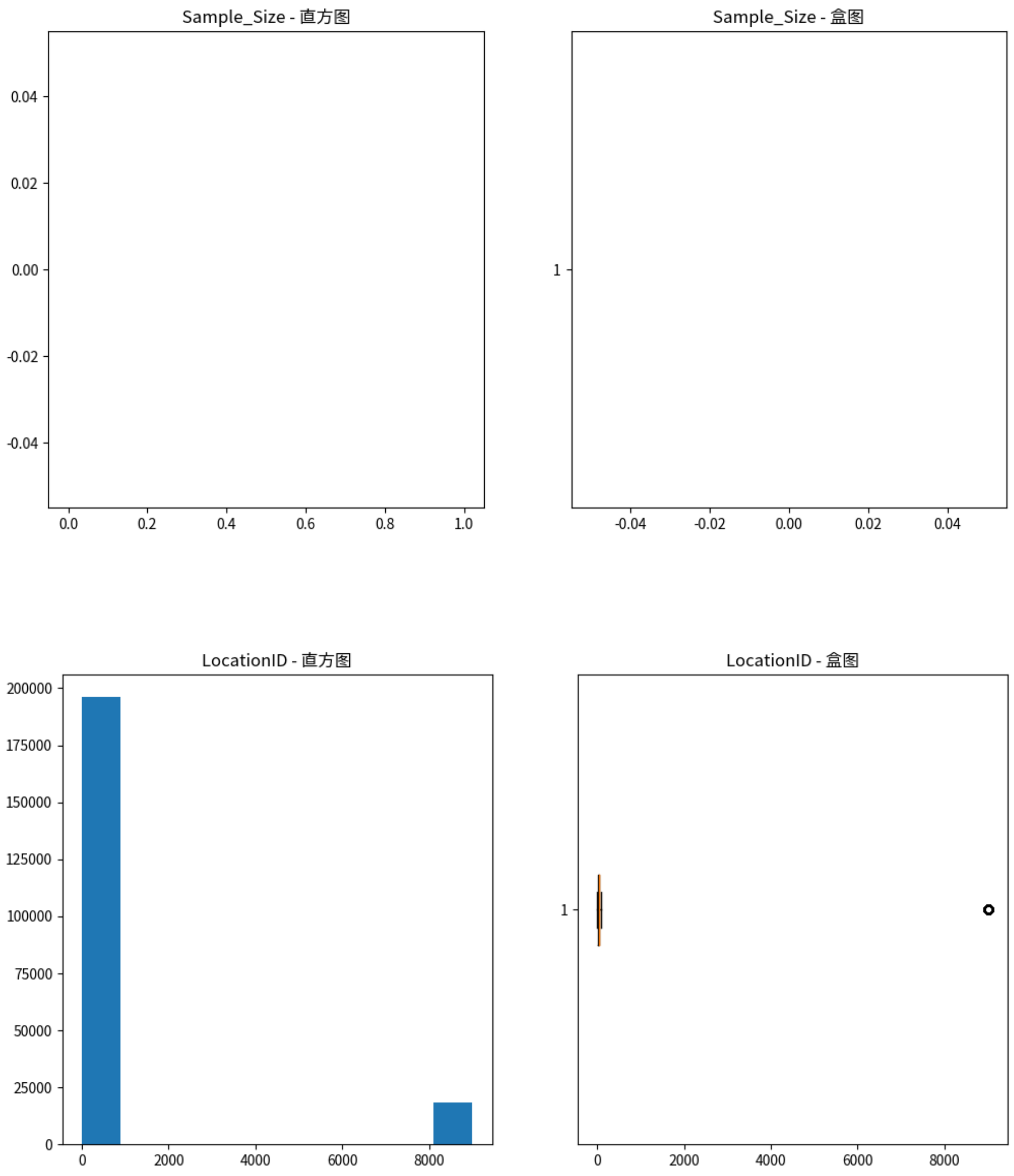
```
from pathlib import Path
import matplotlib as mpl
from matplotlib import font_manager
font_path = Path('/usr/share/fonts/opentype/noto')

#   matplotlib
mpl.rcParams['font.family'] = font_manager.FontProperties(fname="/usr/share/
 ↪fonts/opentype/noto/NotoSansCJK-Regular.ttc").get_name()
mpl.rcParams['axes.unicode_minus'] = False

#     -
for column in num_attributes.columns:
    plt.figure(figsize=(12, 6))

    plt.subplot(121)
    plt.hist(df[column].dropna())
    plt.title(f"{column} -   ")

    plt.subplot(122)
    plt.boxplot(df[column].dropna(), vert=False)
    plt.title(f"{column} -  ")

    plt.show()
```


YearStart - 直方图
40000
30000
20000
10000
0
2015
2016
2017
2018
2019
2020
YearStart - 盒图
1
2015
2016
2017
2018
2019
2020


YearEnd - 直方图
40000
30000
20000
10000
0
2015
2016
2017
2018
2019
2020
YearEnd - 盒图
1
2015
2016
2017
2018
2019
2020

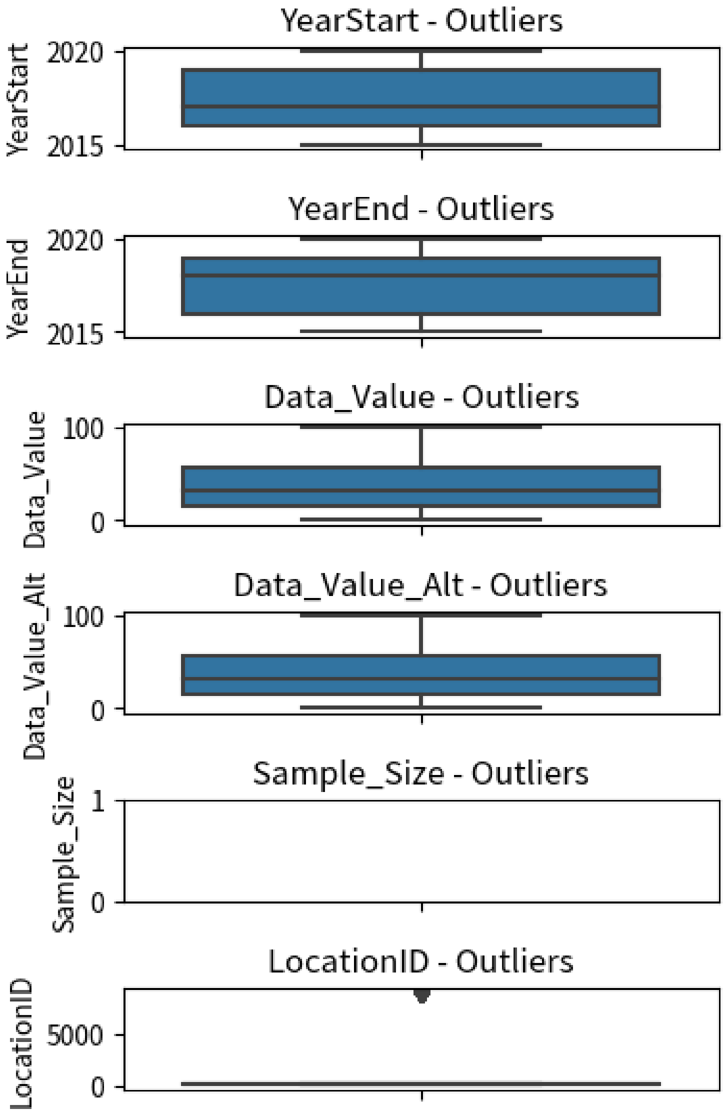

Data_Value - 直方图
25000
20000
15000
10000
5000
0
0
20
40
60
80
100
Data_Value - 盒图
1
0
20
40
60
80
100


Data_Value_Alt - 直方图
25000
20000
15000
10000
5000
0
0
20
40
60
80
100
Data_Value_Alt - 盒图
1
0
20
40
60
80
100



```
[ ]: import seaborn as sns

     #
     numeric_columns = df.select_dtypes(include=['number']).columns

     #         boxplot
     n_plots = len(numeric_columns)
     fig, axes = plt.subplots(nrows=n_plots, ncols=1, figsize=(4, 1 * n_plots))
```

```
for idx, column in enumerate(numeric_columns):
    sns.boxplot(y=df[column], ax=axes[idx])
    axes[idx].set_title(f'{column} - Outliers')
    axes[idx].set_ylabel(column)

#
plt.tight_layout()
plt.show()
```



[ ]:

```
#
missing_values = df.isnull().sum()
```

```
print(missing_values)
```

```
YearStart                         0
YearEnd                           0
LocationAbbr                      0
LocationDesc                      0
Datasource                        0
Class                             0
Topic                             0
Question                          0
Data_Value_Unit                   0
DataValueTypeID                   0
Data_Value_Type                   0
Data_Value                    69833
Data_Value_Alt                69833
Low_Confidence_Limit          69990
High_Confidence_Limit         69990
Sample_Size                  214462
StratificationCategory1           0
Stratification1                   0
StratificationCategory2       27669
Stratification2               27669
Geolocation                   23049
ClassID                           0
TopicID                           0
QuestionID                        0
LocationID                        0
StratificationCategoryID1         0
StratificationID1                 0
StratificationCategoryID2         0
StratificationID2                 0
dtype: int64
```

[ ]:

```
# uncommon  strategy
#

# strategy="del"
# strategy="max_fre"
# strategy="atri_rela"  #
strategy="sim"    #
if strategy=="del":
    #
    df_dropna = df.dropna()
    print(df_dropna.describe())
    print(df.compare(df_dropna))
elif strategy=="max_fre":
    #
    df_filled_max_frequency = df.copy()
    for column in df_filled_max_frequency.columns:
        column_mode = df_filled_max_frequency[column].mode()
        if not column_mode.empty:
            most_frequent_value = column_mode.iloc[0]
            df_filled_max_frequency[column] = df_filled_max_frequency[column].
 ↪fillna(most_frequent_value)
    print(df_filled_max_frequency.describe())
    print(df.compare(df_filled_max_frequency))
elif strategy=="atri_rela":
    from sklearn.linear_model import LinearRegression
    #
    df_regression = df.dropna(subset=['YearStart', 'Data_Value'])
    #
    X = df_regression['YearStart'].values.reshape(-1, 1)
    y = df_regression['Data_Value']
    model = LinearRegression().fit(X, y)

    #
    missing_rows = df['Data_Value'].isnull()
    missing_data_values = df.loc[missing_rows, 'YearStart'].values.reshape(-1,␣
 ↪1)
    predicted_values = model.predict(missing_data_values)
    df_filled_regression = df.copy()
    df_filled_regression.loc[missing_rows, 'Data_Value'] = predicted_values
    print(df_filled_regression.describe())
    print(df.compare(df_filled_regression))
    pass
elif strategy=="sim":
    from sklearn.experimental import enable_iterative_imputer
    from sklearn.impute import SimpleImputer, KNNImputer
    #
    mapping_dict = {}

    #           SimpleImputer
    for col in df:
        df[col] = df[col].astype('category')
        mapping_dict[col] = dict(enumerate(df[col].cat.categories))
        df[col] = df[col].cat.codes

    #  SimpleImputer
    # numerical_imputer = SimpleImputer(strategy='mean')
    # imputed_data = numerical_imputer.fit_transform(df)
    # df_filled = pd.DataFrame(imputed_data, columns=df.columns)
    #  KNNImputer
    knn_imputer = KNNImputer(n_neighbors=5)
    imputed_data = knn_imputer.fit_transform(df)
    df_filled = pd.DataFrame(imputed_data, columns=df.columns)
    #
    for col in mapping_dict:
        reversed_mapping = {v: k for k, v in mapping_dict[col].items()}
        df_filled[col] = df_filled[col].round(0).astype('int').
↪map(reversed_mapping).astype('category')
    print(df_filled.describe())
    print(df.compare(df_filled))
    pass
```

```
        YearStart  YearEnd  LocationAbbr  LocationDesc  Datasource   Class  \
count      214462   214462        214462        214462      214462  214462
unique          6        6            59            59           1       7
top             0        5            49            49           0       4
freq        45980    46966          4644          4644      214462   71694

         Topic  Question  Data_Value_Unit  DataValueTypeID  ...  \
count   214462    214462           214462           214462  ...
unique      39        39                2                2  ...
top         19        14                0                1  ...
freq      8300      8300           197929           197929  ...

        Stratification2  Geolocation  ClassID  TopicID  QuestionID  \
count            214462       214462   214462   214462      214462
unique                8           55        7       39          39
top                   0            0        0        0          14
freq              27669        23049    71694     8300        8300

        LocationID  StratificationCategoryID1  StratificationID1  \
count       214462                     214462             214462
unique          59                          1                  3
top             51                          0                  2
freq          4644                     214462              71919

        StratificationCategoryID2  StratificationID2
count                      214462             214462
unique                          3                  8
top                             2                  6
freq                       134959              27669

[4 rows x 29 columns]
Empty DataFrame
Columns: []
Index: []
```