Healthcare_Insurance_Analysis Project (1)

July 30, 2024

[ ]:
```
# Importing libraries.
```

[ ]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

[ ]:
```
import warnings
warnings.filterwarnings('ignore')
```

[ ]:
```
# Loading the dataset.
```

[ ]:
```
df1 = pd.read_csv(r"C:
 ↪\Users\Admin\Downloads\Healthcare_Insurance_Analysis_Datasets␣
 ↪(2)\Hospitalisation details.csv")
df2 = pd.read_csv(r"C:
 ↪\Users\Admin\Downloads\Healthcare_Insurance_Analysis_Datasets (2)\Medical␣
 ↪Examinations.csv")
```

[ ]:
```
import os
os.getcwd()
```

[ ]:
```
'C:\\Users\\Admin'
```

[ ]:
```
df3 = pd.read_excel(r"C:\Users\Admin\Healthcare_Insurance_Analysis_Datasets␣
 ↪(2)\Names.xlsx")
```

[ ]:
```
# Merging the datasets.
```

[ ]:
```
df4 = pd.merge(left=df1,right=df2,on='Customer ID',how='inner')
```

[ ]:
```
df4
```

[ ]:

	Customer ID	year	month	date	children	charges	Hospital tier	\
0	Id2335	1992	Jul	9	0	563.84	tier - 2	
1	Id2334	1992	Nov	30	0	570.62	tier - 2	
2	Id2333	1993	Jun	30	0	600.00	tier - 2	
3	Id2332	1992	Sep	13	0	604.54	tier - 3	

```
4          Id2331  1998   Jul    27          0    637.26      tier - 3
…            …     …      …     …          …        …           …
2330         Id5   1989   Jun    19          0  55135.40      tier - 1
2331         Id4   1991   Jun     6          1  58571.07      tier - 1
2332         Id3   1970     ?    11          3  60021.40      tier - 1
2333         Id2   1977   Jun     8          0  62592.87      tier - 2
2334         Id1   1968   Oct    12          0  63770.43      tier - 1

     City tier State ID     BMI  HBA1C Heart Issues Any Transplants  \
0     tier - 3    R1013  17.580   4.51           No               No
1     tier - 1    R1013  17.600   4.39           No               No
2     tier - 1    R1013  16.470   6.35           No               No
3     tier - 3    R1013  17.700   6.28           No               No
4     tier - 3    R1013  22.340   5.57           No               No
…        …       …      …     …              …                …
2330  tier - 2    R1012  35.530   5.45           No               No
2331  tier - 3    R1024  38.095   6.05           No               No
2332  tier - 1    R1012  34.485  11.87          yes               No
2333  tier - 3    R1013  30.360   5.77           No               No
2334  tier - 3    R1013  47.410   7.47           No               No

     Cancer history NumberOfMajorSurgeries smoker
0                No                      1     No
1                No                      1     No
2               Yes                      1     No
3                No                      1     No
4                No                      1     No
…              …                       …    …
2330             No       No major surgery    yes
2331             No       No major surgery    yes
2332             No                      2    yes
2333             No       No major surgery    yes
2334             No       No major surgery    yes

[2335 rows x 16 columns]
```

[ ]:
```
df = pd.merge(left=df3,right=df4,on='Customer ID',how='inner')
```

[ ]:
```
df
```

[ ]:
```
     Customer ID                                name  year month  date  \
0            Id1                   Hawks, Ms.  Kelly  1968   Oct    12
1            Id2               Lehner, Mr.  Matthew D  1977   Jun     8
2            Id3                        Lu, Mr.  Phil  1970     ?    11
3            Id4                 Osborne, Ms.  Kelsey  1991   Jun     6
4            Id5                 Kadala, Ms.  Kristyn  1989   Jun    19
…          …                                    …   …   …   …
```

```
2330      Id2331                Brietzke, Mr.  Jordan  1998   Jul    27
2331      Id2332  Riveros Gonzalez, Mr.  Juan D. Sr.  1992   Sep    13
2332      Id2333                   Albano, Ms.  Julie  1993   Jun    30
2333      Id2334               Rosendahl, Mr.  Evan P  1992   Nov    30
2334      Id2335                 German, Mr.  Aaron K  1992   Jul     9

      children   charges Hospital tier City tier State ID     BMI  HBA1C  \
0            0  63770.43      tier - 1  tier - 3    R1013  47.410   7.47
1            0  62592.87      tier - 2  tier - 3    R1013  30.360   5.77
2            3  60021.40      tier - 1  tier - 1    R1012  34.485  11.87
3            1  58571.07      tier - 1  tier - 3    R1024  38.095   6.05
4            0  55135.40      tier - 1  tier - 2    R1012  35.530   5.45
…          …       …            …         …       …     …     …
2330         0    637.26      tier - 3  tier - 3    R1013  22.340   5.57
2331         0    604.54      tier - 3  tier - 3    R1013  17.700   6.28
2332         0    600.00      tier - 2  tier - 1    R1013  16.470   6.35
2333         0    570.62      tier - 2  tier - 1    R1013  17.600   4.39
2334         0    563.84      tier - 2  tier - 3    R1013  17.580   4.51

     Heart Issues Any Transplants Cancer history NumberOfMajorSurgeries smoker
0              No               No              No       No major surgery    yes
1              No               No              No       No major surgery    yes
2             yes               No              No                      2    yes
3              No               No              No       No major surgery    yes
4              No               No              No       No major surgery    yes
…            …              …             …                    …    …
2330           No               No              No                      1     No
2331           No               No              No                      1     No
2332           No               No             Yes                      1     No
2333           No               No              No                      1     No
2334           No               No              No                      1     No

[2335 rows x 17 columns]
```

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

[ ]:
```
  Customer ID                   name  year month  date  children   charges  \
0         Id1       Hawks, Ms.  Kelly  1968   Oct    12         0  63770.43
1         Id2  Lehner, Mr.  Matthew D  1977   Jun     8         0  62592.87
2         Id3           Lu, Mr.  Phil  1970     ?    11         3  60021.40
3         Id4    Osborne, Ms.  Kelsey  1991   Jun     6         1  58571.07
4         Id5    Kadala, Ms.  Kristyn  1989   Jun    19         0  55135.40

  Hospital tier City tier State ID     BMI  HBA1C Heart Issues  \
0      tier - 1  tier - 3    R1013  47.410   7.47           No
1      tier - 2  tier - 3    R1013  30.360   5.77           No
2      tier - 1  tier - 1    R1012  34.485  11.87          yes
```

```
3      tier - 1  tier - 3    R1024  38.095   6.05           No
4      tier - 1  tier - 2    R1012  35.530   5.45           No

  Any Transplants Cancer history NumberOfMajorSurgeries smoker
0              No             No       No major surgery    yes
1              No             No       No major surgery    yes
2              No             No                       2    yes
3              No             No       No major surgery    yes
4              No             No       No major surgery    yes
```

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2335 entries, 0 to 2334
Data columns (total 17 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Customer ID             2335 non-null   object
 1   name                    2335 non-null   object
 2   year                    2335 non-null   object
 3   month                   2335 non-null   object
 4   date                    2335 non-null   int64
 5   children                2335 non-null   int64
 6   charges                 2335 non-null   float64
 7   Hospital tier           2335 non-null   object
 8   City tier               2335 non-null   object
 9   State ID                2335 non-null   object
 10  BMI                     2335 non-null   float64
 11  HBA1C                   2335 non-null   float64
 12  Heart Issues            2335 non-null   object
 13  Any Transplants         2335 non-null   object
 14  Cancer history          2335 non-null   object
 15  NumberOfMajorSurgeries  2335 non-null   object
 16  smoker                  2335 non-null   object
dtypes: float64(3), int64(2), object(12)
memory usage: 310.2+ KB
```

[ ]:
```
# Finding Missing Values.
```

[ ]:
```
df.isna().sum()
```

[ ]:
```
Customer ID             0
name                    0
year                    0
month                   0
date                    0
children                0
charges                 0
```

```
Hospital tier             0
City tier                 0
State ID                  0
BMI                       0
HBA1C                     0
Heart Issues              0
Any Transplants           0
Cancer history            0
NumberOfMajorSurgeries    0
smoker                    0
dtype: int64
```

[ ]:
```
# Finding trivial rows and its percentage.
```

[ ]:
```
trivial_rows = df[df== "?"].count(axis=1).sum()
```

[ ]:
```
trivial_rows
```

[ ]: 11

[ ]:
```
 total_rows = df.shape[0]
total_rows
```

[ ]: 2335

[ ]:
```
percentage = (trivial_rows/total_rows)*100
percentage
```

[ ]: 0.47109207708779444

[ ]:
```
print("The Percentage of trivial rows: {:.2f}%".format(percentage))
```

```
The Percentage of trivial rows: 0.47%
```

[ ]:
```
df = df[df != "?"].dropna()
```

[ ]:
```
df.shape
```

[ ]: (2325, 17)

[ ]:
```
df.describe()
```

[ ]:

	date	children	charges	BMI	HBA1C
count	2325.000000	2325.000000	2325.000000	2325.000000	2325.000000
mean	15.572903	1.025376	13521.660254	30.995630	6.576718
std	8.720287	1.234456	11863.492697	8.744938	2.226892
min	1.000000	0.000000	563.840000	15.010000	4.000000
25%	8.000000	0.000000	5116.500000	24.605000	4.900000
50%	15.000000	0.000000	9634.540000	30.400000	5.810000

```
75%       23.000000      2.000000  16903.500000     36.300000      7.950000
max       30.000000      5.000000  63770.430000     55.050000     12.000000
```

```
[ ]: # Handling nominal and ordinal categories.
```

```
[ ]: from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
```

```
[ ]: df['Heart Issues'].value_counts()
```

```
[ ]: Heart Issues
No     1405
yes     920
Name: count, dtype: int64
```

```
[ ]: df['Any Transplants'].value_counts()
```

```
[ ]: Any Transplants
No     2183
yes     142
Name: count, dtype: int64
```

```
[ ]: df['Cancer history'].value_counts()
```

```
[ ]: Cancer history
No     1934
Yes     391
Name: count, dtype: int64
```

```
[ ]: df['Heart Issues'] = le.fit_transform(df['Heart Issues'])
df['Any Transplants'] = le.fit_transform(df['Any Transplants'])
df['Cancer history'] = le.fit_transform(df['Cancer history'])
df['smoker'] = le.fit_transform(df['smoker'])
```

```
[ ]: df['Heart Issues'].value_counts()
```

```
[ ]: Heart Issues
0    1405
1     920
Name: count, dtype: int64
```

```
[ ]: df['Any Transplants'].value_counts()
```

```
[ ]: Any Transplants
0    2183
1     142
Name: count, dtype: int64
```

```
[ ]: df['Cancer history'].value_counts()
```

```
[ ]: Cancer history
     0    1934
     1     391
     Name: count, dtype: int64
```

```
[ ]: df['Hospital tier'].value_counts()
```

```
[ ]: Hospital tier
     tier - 2    1334
     tier - 3     691
     tier - 1     300
     Name: count, dtype: int64
```

```
[ ]: df['smoker'].value_counts()
```

```
[ ]: smoker
     0    1839
     1     486
     Name: count, dtype: int64
```

```
[ ]: df['Hospital tier'] = df['Hospital tier'].str.replace('tier -','')
```

```
[ ]: df['Hospital tier'].value_counts()
```

```
[ ]: Hospital tier
     2    1334
     3     691
     1     300
     Name: count, dtype: int64
```

```
[ ]: df['City tier'].value_counts()
```

```
[ ]: City tier
     tier - 2    807
     tier - 3    789
     tier - 1    729
     Name: count, dtype: int64
```

```
[ ]: df['City tier'] = df['City tier'].str.replace('tier -','')
```

```
[ ]: df['City tier'].value_counts()
```

```
[ ]: City tier
     2    807
     3    789
     1    729
```

```
Name: count, dtype: int64
```

[ ]:
```
df
```

[ ]:
```
     Customer ID                                name  year month  date  \
0            Id1                   Hawks, Ms.  Kelly  1968   Oct    12
1            Id2               Lehner, Mr.  Matthew D  1977   Jun     8
3            Id4                 Osborne, Ms.  Kelsey  1991   Jun     6
4            Id5                 Kadala, Ms.  Kristyn  1989   Jun    19
5            Id6               Baker, Mr.  Russell B.  1962   Aug     4
…            …                                    …    …    …    …
2330      Id2331                Brietzke, Mr.  Jordan  1998   Jul    27
2331      Id2332  Riveros Gonzalez, Mr.  Juan D. Sr.  1992   Sep    13
2332      Id2333                   Albano, Ms.  Julie  1993   Jun    30
2333      Id2334               Rosendahl, Mr.  Evan P  1992   Nov    30
2334      Id2335                 German, Mr.  Aaron K  1992   Jul     9

      children   charges Hospital tier City tier State ID     BMI  HBA1C  \
0            0  63770.43             1         3    R1013  47.410   7.47
1            0  62592.87             2         3    R1013  30.360   5.77
3            1  58571.07             1         3    R1024  38.095   6.05
4            0  55135.40             1         2    R1012  35.530   5.45
5            0  52590.83             1         3    R1011  32.800   6.59
…          …       …              …         …      …     …    …
2330         0    637.26             3         3    R1013  22.340   5.57
2331         0    604.54             3         3    R1013  17.700   6.28
2332         0    600.00             2         1    R1013  16.470   6.35
2333         0    570.62             2         1    R1013  17.600   4.39
2334         0    563.84             2         3    R1013  17.580   4.51

      Heart Issues  Any Transplants  Cancer history NumberOfMajorSurgeries  \
0                0                0               0       No major surgery
1                0                0               0       No major surgery
3                0                0               0       No major surgery
4                0                0               0       No major surgery
5                0                0               0       No major surgery
…              …                …               …                       …
2330             0                0               0                       1
2331             0                0               0                       1
2332             0                0               1                       1
2333             0                0               0                       1
2334             0                0               0                       1

      smoker
0          1
1          1
3          1
```

```
4          1
5          1
…        …
2330       0
2331       0
2332       0
2333       0
2334       0

[2325 rows x 17 columns]
```

[ ]:
```
# Filtering Based on Required STATE ID.
```

[ ]:
```
df['State ID'].value_counts()
```

[ ]:
```
State ID
R1013    609
R1011    574
R1012    572
R1024    159
R1026     84
R1021     70
R1016     64
R1025     40
R1023     38
R1017     36
R1019     26
R1022     14
R1014     13
R1015     11
R1018      9
R1020      6
Name: count, dtype: int64
```

[ ]:
```
df['state_group'] = np.where((df['State ID'] == 'R1011') | (df['State ID'] ==␣
↪'R1012') | (df['State ID'] == 'R1013'), df['State ID'], 'other')
```

[ ]:
```
df
```

[ ]:
```
     Customer ID                              name  year month  date  \
0            Id1                 Hawks, Ms.  Kelly  1968   Oct    12
1            Id2             Lehner, Mr.  Matthew D  1977   Jun     8
3            Id4               Osborne, Ms.  Kelsey  1991   Jun     6
4            Id5               Kadala, Ms.  Kristyn  1989   Jun    19
5            Id6             Baker, Mr.  Russell B.  1962   Aug     4
…          …                                  …    …    …    …
2330      Id2331              Brietzke, Mr.  Jordan  1998   Jul    27
```

```
2331      Id2332  Riveros Gonzalez, Mr.  Juan D. Sr.  1992   Sep    13
2332      Id2333                  Albano, Ms.  Julie  1993   Jun    30
2333      Id2334              Rosendahl, Mr.  Evan P  1992   Nov    30
2334      Id2335                German, Mr.  Aaron K  1992   Jul     9

      children   charges Hospital tier City tier State ID     BMI  HBA1C  \
0            0  63770.43             1         3    R1013  47.410   7.47
1            0  62592.87             2         3    R1013  30.360   5.77
3            1  58571.07             1         3    R1024  38.095   6.05
4            0  55135.40             1         2    R1012  35.530   5.45
5            0  52590.83             1         3    R1011  32.800   6.59
…        …       …            …        …       …     …    …
2330         0    637.26             3         3    R1013  22.340   5.57
2331         0    604.54             3         3    R1013  17.700   6.28
2332         0    600.00             2         1    R1013  16.470   6.35
2333         0    570.62             2         1    R1013  17.600   4.39
2334         0    563.84             2         3    R1013  17.580   4.51

      Heart Issues  Any Transplants  Cancer history NumberOfMajorSurgeries  \
0                0                0               0       No major surgery
1                0                0               0       No major surgery
3                0                0               0       No major surgery
4                0                0               0       No major surgery
5                0                0               0       No major surgery
…              …                …               …                    …
2330             0                0               0                    1
2331             0                0               0                    1
2332             0                0               1                    1
2333             0                0               0                    1
2334             0                0               0                    1

      smoker state_group
0          1       R1013
1          1       R1013
3          1       other
4          1       R1012
5          1       R1011
…      …          …
2330       0       R1013
2331       0       R1013
2332       0       R1013
2333       0       R1013
2334       0       R1013

[2325 rows x 18 columns]
```

[ ]:
```
df=df[df["State ID"].isin(['R1011','R1012','R1013'])]
df.shape
df["State ID"]=le.fit_transform(df["State ID"])
df["State ID"].unique()
```

[ ]:
```
array([2, 1, 0])
```

[ ]:
```
df['state_group'].value_counts()
```

[ ]:
```
state_group
R1013    609
R1011    574
R1012    572
Name: count, dtype: int64
```

[ ]:
```
df["state_group"].replace('R1011',1,inplace=True)
df["state_group"].replace('R1012',2,inplace=True)
df["state_group"].replace('R1013',3,inplace=True)
df["state_group"].replace('other',0,inplace=True)
```

[ ]:
```
df
```

[ ]:

	Customer ID	name	year	month	date	\
0	Id1	Hawks, Ms. Kelly	1968	Oct	12	
1	Id2	Lehner, Mr. Matthew D	1977	Jun	8	
4	Id5	Kadala, Ms. Kristyn	1989	Jun	19	
5	Id6	Baker, Mr. Russell B.	1962	Aug	4	
6	Id7	Macpherson, Mr. Scott	1994	Oct	27	
…	…	…	…	…	…	
2330	Id2331	Brietzke, Mr. Jordan	1998	Jul	27	
2331	Id2332	Riveros Gonzalez, Mr. Juan D. Sr.	1992	Sep	13	
2332	Id2333	Albano, Ms. Julie	1993	Jun	30	
2333	Id2334	Rosendahl, Mr. Evan P	1992	Nov	30	
2334	Id2335	German, Mr. Aaron K	1992	Jul	9	

	children	charges	Hospital tier	City tier	State ID	BMI	HBA1C	\
0	0	63770.43	1	3	2	47.41	7.47	
1	0	62592.87	2	3	2	30.36	5.77	
4	0	55135.40	1	2	1	35.53	5.45	
5	0	52590.83	1	3	0	32.80	6.59	
6	1	51194.56	1	3	0	36.40	6.07	
…	…	…	…	…	…	…	…	
2330	0	637.26	3	3	2	22.34	5.57	
2331	0	604.54	3	3	2	17.70	6.28	
2332	0	600.00	2	1	2	16.47	6.35	
2333	0	570.62	2	1	2	17.60	4.39	
2334	0	563.84	2	3	2	17.58	4.51	

```
      Heart Issues  Any Transplants  Cancer history NumberOfMajorSurgeries  \
0                0                0               0       No major surgery
1                0                0               0       No major surgery
4                0                0               0       No major surgery
5                0                0               0       No major surgery
6                0                0               0       No major surgery
…              …                …               …                       …
2330             0                0               0                       1
2331             0                0               0                       1
2332             0                0               1                       1
2333             0                0               0                       1
2334             0                0               0                       1

      smoker  state_group
0          1            3
1          1            3
4          1            2
5          1            1
6          1            1
…        …            …
2330       0            3
2331       0            3
2332       0            3
2333       0            3
2334       0            3

[1755 rows x 18 columns]
```

[ ]:
```
# Handling "Number of Major Surgeries" column.
```

[ ]:
```
df['NumberOfMajorSurgeries'].replace('No major surgery','0',inplace = True)
```

[ ]:
```
df.tail(10)
```

[ ]:
```
     Customer ID                                name  year month  date  \
2325      Id2326             Castro, Mr.  Sebastian  1997   Nov     9
2326      Id2327                 Howell, Ms.  Laura  2002   Nov    29
2327      Id2328                 Avery, Ms.  Nicole  1995   Jul     4
2328      Id2329             Bohinski, Ms.  Susan E  1993   Jun     1
2329      Id2330                   Kohls, Ms.  Katy  2001   Nov    20
2330      Id2331              Brietzke, Mr.  Jordan  1998   Jul    27
2331      Id2332  Riveros Gonzalez, Mr.  Juan D. Sr.  1992   Sep    13
2332      Id2333                 Albano, Ms.  Julie  1993   Jun    30
2333      Id2334             Rosendahl, Mr.  Evan P  1992   Nov    30
2334      Id2335               German, Mr.  Aaron K  1992   Jul     9
```

```
      children  charges Hospital tier City tier  State ID    BMI  HBA1C  \
2325         0   670.00             3         3         2  20.10   5.60
2326         0   668.00             3         2         1  21.77  10.67
2327         0   650.00             3         3         2  17.82   5.26
2328         0   650.00             3         3         2  17.07   5.22
2329         0   646.14             3         3         1  22.24   4.29
2330         0   637.26             3         3         2  22.34   5.57
2331         0   604.54             3         3         2  17.70   6.28
2332         0   600.00             2         1         2  16.47   6.35
2333         0   570.62             2         1         2  17.60   4.39
2334         0   563.84             2         3         2  17.58   4.51

      Heart Issues  Any Transplants  Cancer history NumberOfMajorSurgeries  \
2325             1                0               1                       1
2326             0                0               0                       0
2327             1                0               0                       1
2328             0                0               1                       1
2329             1                0               0                       0
2330             0                0               0                       1
2331             0                0               0                       1
2332             0                0               1                       1
2333             0                0               0                       1
2334             0                0               0                       1

      smoker  state_group
2325       0            3
2326       0            2
2327       0            3
2328       0            3
2329       0            2
2330       0            3
2331       0            3
2332       0            3
2333       0            3
2334       0            3
```

[ ]:
```
# Calculating Age.
```

[ ]:
```
df['year'] = df.year.astype(int)
```

[ ]:
```
df.dtypes
```

[ ]:
```
Customer ID               object
name                      object
year                       int32
month                     object
date                       int64
```

```
children                    int64
charges                   float64
Hospital tier              object
City tier                  object
State ID                    int32
BMI                       float64
HBA1C                     float64
Heart Issues                int32
Any Transplants             int32
Cancer history              int32
NumberOfMajorSurgeries     object
smoker                      int32
state_group                 int64
dtype: object
```

[ ]:
```
df['Age'] = 2024 - df['year']
```

[ ]:
```
df
```

[ ]:
```
     Customer ID                                name  year month  date  \
0            Id1                   Hawks, Ms.  Kelly  1968   Oct    12
1            Id2               Lehner, Mr.  Matthew D  1977   Jun     8
4            Id5                 Kadala, Ms.  Kristyn  1989   Jun    19
5            Id6               Baker, Mr.  Russell B.  1962   Aug     4
6            Id7               Macpherson, Mr.  Scott  1994   Oct    27
…            …                                     …    …    …    …
2330      Id2331                Brietzke, Mr.  Jordan  1998   Jul    27
2331      Id2332  Riveros Gonzalez, Mr.  Juan D. Sr.  1992   Sep    13
2332      Id2333                   Albano, Ms.  Julie  1993   Jun    30
2333      Id2334               Rosendahl, Mr.  Evan P  1992   Nov    30
2334      Id2335                 German, Mr.  Aaron K  1992   Jul     9

      children   charges Hospital tier City tier  State ID    BMI  HBA1C  \
0            0  63770.43             1         3         2  47.41   7.47
1            0  62592.87             2         3         2  30.36   5.77
4            0  55135.40             1         2         1  35.53   5.45
5            0  52590.83             1         3         0  32.80   6.59
6            1  51194.56             1         3         0  36.40   6.07
…        …       …            …        …        …    …    …
2330         0    637.26             3         3         2  22.34   5.57
2331         0    604.54             3         3         2  17.70   6.28
2332         0    600.00             2         1         2  16.47   6.35
2333         0    570.62             2         1         2  17.60   4.39
2334         0    563.84             2         3         2  17.58   4.51

      Heart Issues  Any Transplants  Cancer history NumberOfMajorSurgeries  \
0                0                0               0                       0
```

```
1                0                0                0                    0
4                0                0                0                    0
5                0                0                0                    0
6                0                0                0                    0
…              …                …                …                    …
2330             0                0                0                    1
2331             0                0                0                    1
2332             0                0                1                    1
2333             0                0                0                    1
2334             0                0                0                    1

      smoker  state_group  Age
0          1            3   56
1          1            3   47
4          1            2   35
5          1            1   62
6          1            1   30
…        …            …   …
2330       0            3   26
2331       0            3   32
2332       0            3   31
2333       0            3   32
2334       0            3   32

[1755 rows x 19 columns]
```

[ ]:
```
# Calculating Gender From the Given "Name".
```

[ ]:
```
gender = ['0' if 'Mr.' in name else '1' for name in df['name']]
df["Gender"] = gender
```

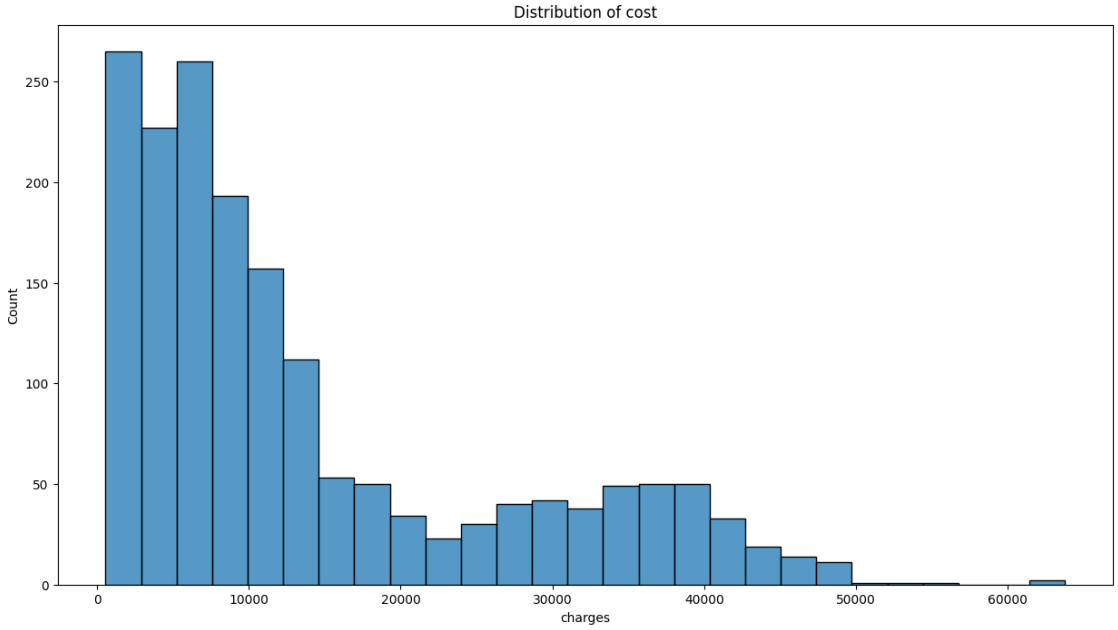
[ ]:
```
df.head()
```

[ ]:
```
  Customer ID                    name  year month  date  children   charges  \
0         Id1       Hawks, Ms.  Kelly  1968   Oct    12         0  63770.43
1         Id2  Lehner, Mr.  Matthew D  1977   Jun     8         0  62592.87
4         Id5    Kadala, Ms.  Kristyn  1989   Jun    19         0  55135.40
5         Id6  Baker, Mr.  Russell B.  1962   Aug     4         0  52590.83
6         Id7  Macpherson, Mr.  Scott  1994   Oct    27         1  51194.56

  Hospital tier City tier  State ID    BMI  HBA1C  Heart Issues  \
0             1         3         2  47.41   7.47             0
1             2         3         2  30.36   5.77             0
4             1         2         1  35.53   5.45             0
5             1         3         0  32.80   6.59             0
6             1         3         0  36.40   6.07             0
```

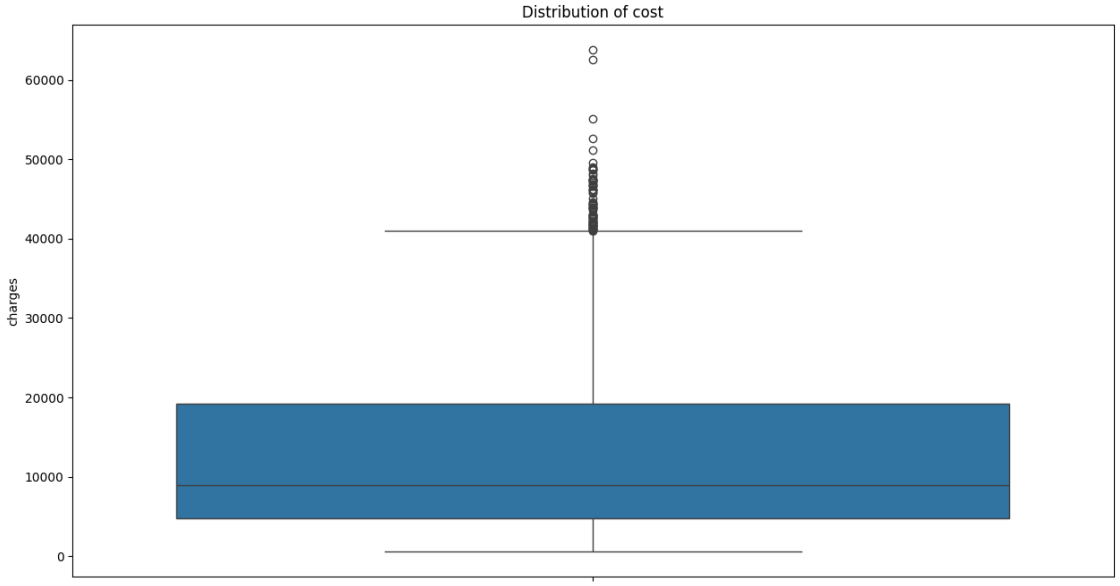
```
   Any Transplants  Cancer history NumberOfMajorSurgeries  smoker  \
0                0               0                       0       1
1                0               0                       0       1
4                0               0                       0       1
5                0               0                       0       1
6                0               0                       0       1

   state_group  Age Gender
0            3   56      1
1            3   47      0
4            2   35      1
5            1   62      0
6            1   30      0
```

[ ]:
```
#   Distribution of costs using a histogram, box and whisker plot, and swarm plot.
```

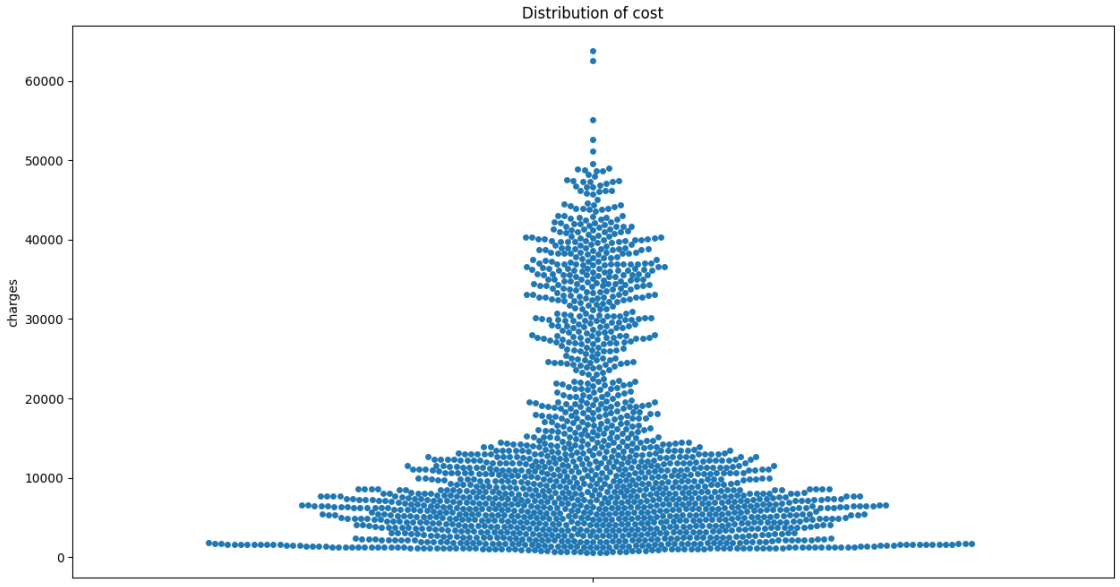
[ ]:
```
plt.figure(figsize=(15,8))
sns.histplot(df['charges'])
plt.title('Distribution of cost')
plt.show()
```



[ ]:
```
plt.figure(figsize=(15,8))
sns.boxplot(df['charges'])
plt.title('Distribution of cost')
plt.show()
```

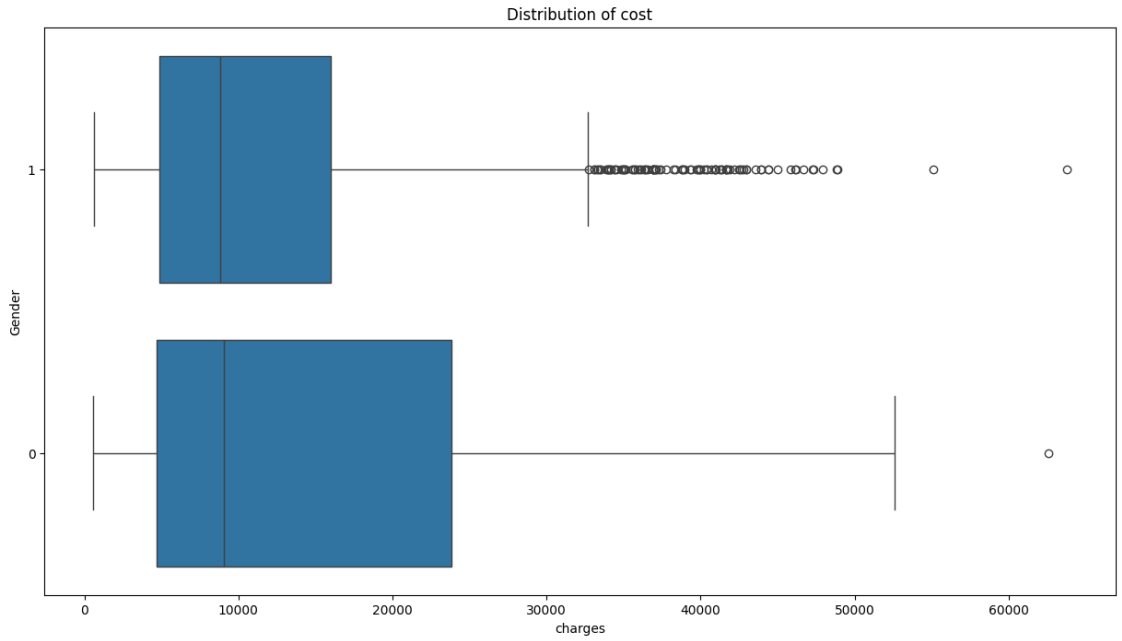


[ ]:
```
plt.figure(figsize=(15,8))
sns.swarmplot(df['charges'])
plt.title('Distribution of cost')
plt.show()
```



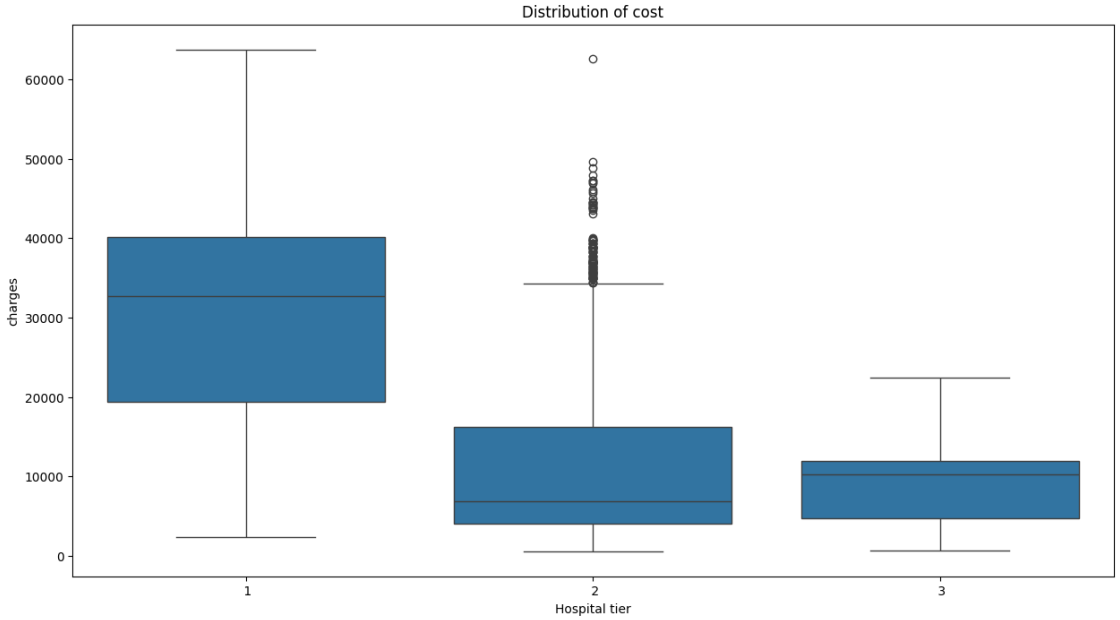
[ ]:
```
# Distribution of charges across gender.
```

[ ]:
```
plt.figure(figsize=(15,8))
sns.boxplot(x = 'charges', y = 'Gender',data = df)
plt.title('Distribution of cost')
plt.show()
```



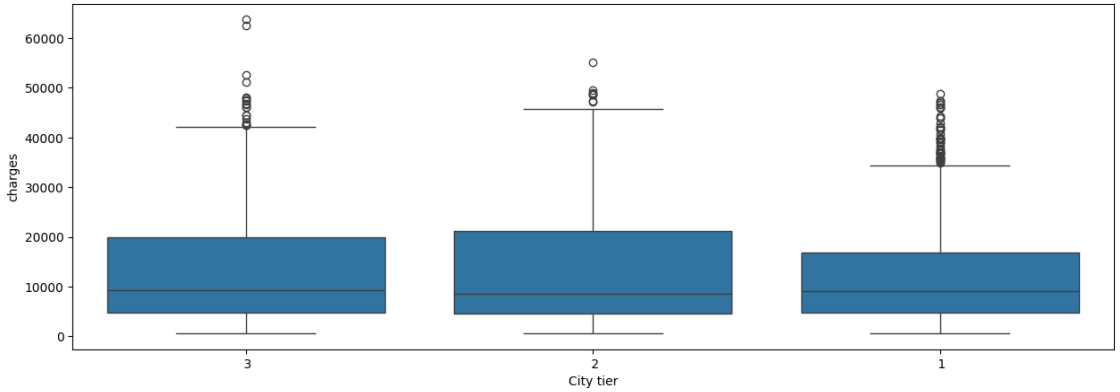
[ ]:
```
#Distribution of charges across Hospital tier.
```

[ ]:
```
plt.figure(figsize=(15,8))
sns.boxplot(x = 'Hospital tier', y = 'charges',data = df)
plt.title('Distribution of cost')
plt.show()
```



[ ]:
```
#Distribution of charges across city tier.
```

[ ]:
```
plt.figure(figsize = (15,5))
sns.boxplot(x = "City tier",y = "charges", data = df)
plt.show()
```

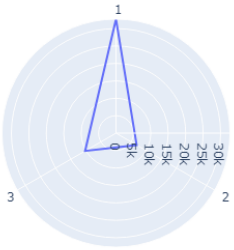


[ ]:
```
# Radar Chart.
```

[ ]:
```
median = df.groupby('Hospital tier')[['charges']].median().reset_index()
```

[ ]:
```
import plotly.express as px
fig = px.line_polar(median, r='charges', theta='Hospital tier', line_close=True)
```

```
fig.show()
```



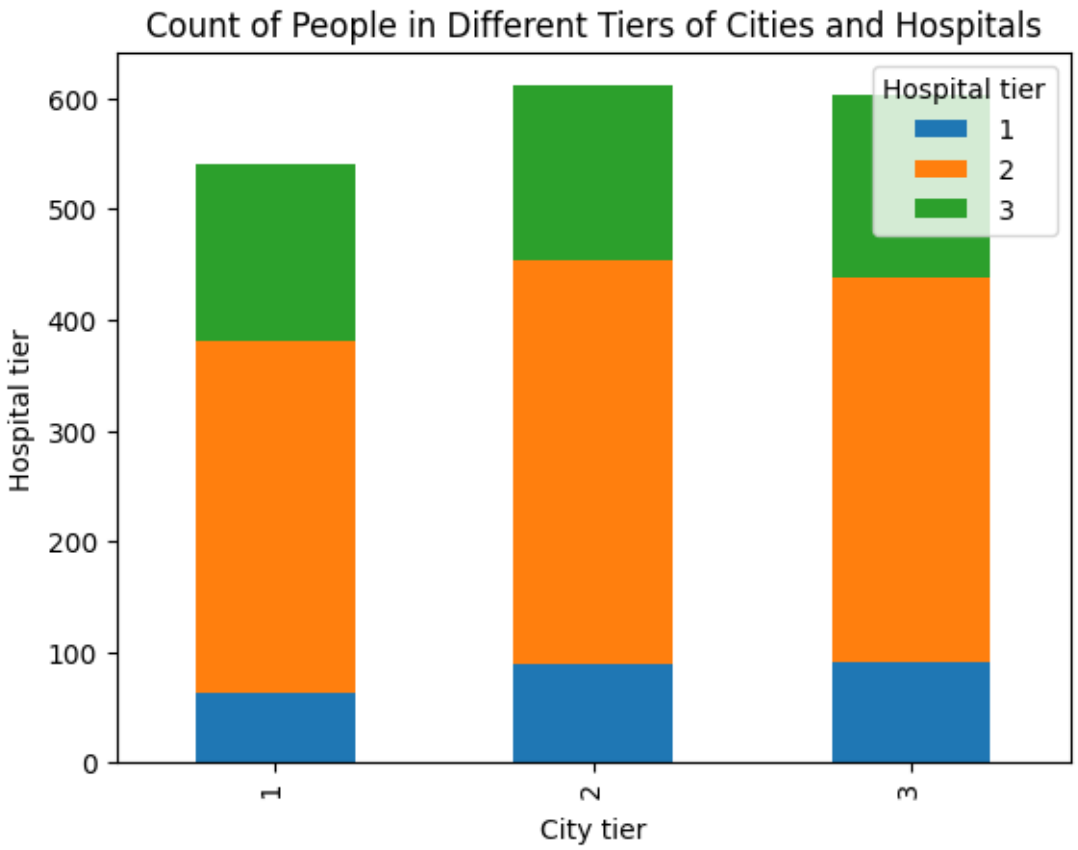
[ ]:
```
# Frequency Table.
```

[ ]:
```
table = pd.crosstab(df['City tier'], df['Hospital tier'])
print(table)
```

```
Hospital tier   1    2    3
City tier
1              64  317  160
2              89  365  157
3              91  348  164
```

[ ]:
```
# Stacked Bar Chart.
```

[ ]:
```
table.plot(kind='bar', stacked=True)
plt.xlabel('City tier')
plt.ylabel('Hospital tier')
plt.title('Count of People in Different Tiers of Cities and Hospitals')
plt.show()
```



[ ]:
```
# Testing Null Hypothesis
```

[ ]:
```
# Average hospitalization costs for the three types of hospitals are not␣
↪significantly different.
```

[ ]:
```
import scipy.stats as stats
print('Null Hypothesis => Average hospitalization costs for the three types of␣
↪hospitals are not significantly different.')
f_val, p_val = stats.f_oneway(df[df['Hospital tier'] == 'tier,1']['charges'],
                              df[df['Hospital tier'] == 'tier,2']['charges'],
                              df[df['Hospital tier'] == 'tier,3']['charges'])
print('P-value :',p_val)
if p_val < 0.05:
    print("Reject null hypothesis")
else:
    print("Accept null hypothesis")
```

```
Null Hypothesis => Average hospitalization costs for the three types of
hospitals are not significantly different.
P-value : nan
```

```
Accept null hypothesis
```

[ ]:
```
# Average hospitalization costs for the three types of cities are not␣
 ↪significantly different.
```

[ ]:
```
print('Null Hypothesis => Average hospitalization costs for the three types of␣
 ↪cities are not significantly different.')
f_val, p_val = stats.f_oneway(df[df['City tier'] == 'tier,1']['charges'],
                              df[df['City tier'] == 'tier,2']['charges'],
                              df[df['City tier'] == 'tier,3']['charges'])
print('P-value :',p_val)
if p_val < 0.05:
    print("Reject null hypothesis")
else:
    print("Accept null hypothesis")
```

```
Null Hypothesis => Average hospitalization costs for the three types of cities
are not significantly different.
P-value : nan
Accept null hypothesis
```

[ ]:
```
# Average hospitalization costs for smokers is not significantly different from␣
 ↪the average cost for nonsmokers.
```

[ ]:
```
print('Null Hypothesis => Average hospitalization costs for smokers is not␣
 ↪significantly different from the average cost for nonsmokers.')
t_val, p_val = stats.ttest_ind(df[df['smoker'] == 'yes']['charges'],
                               df[df['smoker'] == 'no']['charges'])

print('P-value :',p_val)
if p_val < 0.05:
    print("Reject null hypothesis")
else:
    print("Accept null hypothesis")
```

```
Null Hypothesis => Average hospitalization costs for smokers is not
significantly different from the average cost for nonsmokers.
P-value : nan
Accept null hypothesis
```

[ ]:
```
# Smoking and heart issues are independent.
```

[ ]:
```
 from scipy.stats import chi2_contingency
contingency_table = pd.crosstab(df['smoker'], df['Heart Issues'])
chi2, p, dof, expected = chi2_contingency(contingency_table)
print(f'P-value = {p}')
if p < 0.05:
```

```
    print("Reject the null hypothesis, Smoking and heart issues are independent.
  ↪")
else:
    print("Accept null hypothesis, Smoking and heart issues are independent.")
```

```
P-value = 0.9107065371179246
Accept null hypothesis, Smoking and heart issues are independent.
```

[ ]:
```
df.drop(["Customer ID",'name'], inplace=True, axis=1)
df
```

[ ]:

	year	month	date	children	charges	Hospital tier	City tier	State ID	\
0	1968	Oct	12	0	63770.43	1	3	2	
1	1977	Jun	8	0	62592.87	2	3	2	
4	1989	Jun	19	0	55135.40	1	2	1	
5	1962	Aug	4	0	52590.83	1	3	0	
6	1994	Oct	27	1	51194.56	1	3	0	
…	…	…	…	…	…	…	…	…	
2330	1998	Jul	27	0	637.26	3	3	2	
2331	1992	Sep	13	0	604.54	3	3	2	
2332	1993	Jun	30	0	600.00	2	1	2	
2333	1992	Nov	30	0	570.62	2	1	2	
2334	1992	Jul	9	0	563.84	2	3	2	

	BMI	HBA1C	Heart Issues	Any Transplants	Cancer history	\
0	47.41	7.47	0	0	0	
1	30.36	5.77	0	0	0	
4	35.53	5.45	0	0	0	
5	32.80	6.59	0	0	0	
6	36.40	6.07	0	0	0	
…	…	…	…	…	…	
2330	22.34	5.57	0	0	0	
2331	17.70	6.28	0	0	0	
2332	16.47	6.35	0	0	1	
2333	17.60	4.39	0	0	0	
2334	17.58	4.51	0	0	0	

	NumberOfMajorSurgeries	smoker	state_group	Age	Gender
0	0	1	3	56	1
1	0	1	3	47	0
4	0	1	2	35	1
5	0	1	1	62	0
6	0	1	1	30	0
…	…	…	…	…	…
2330	1	0	3	26	0
2331	1	0	3	32	0
2332	1	0	3	31	1

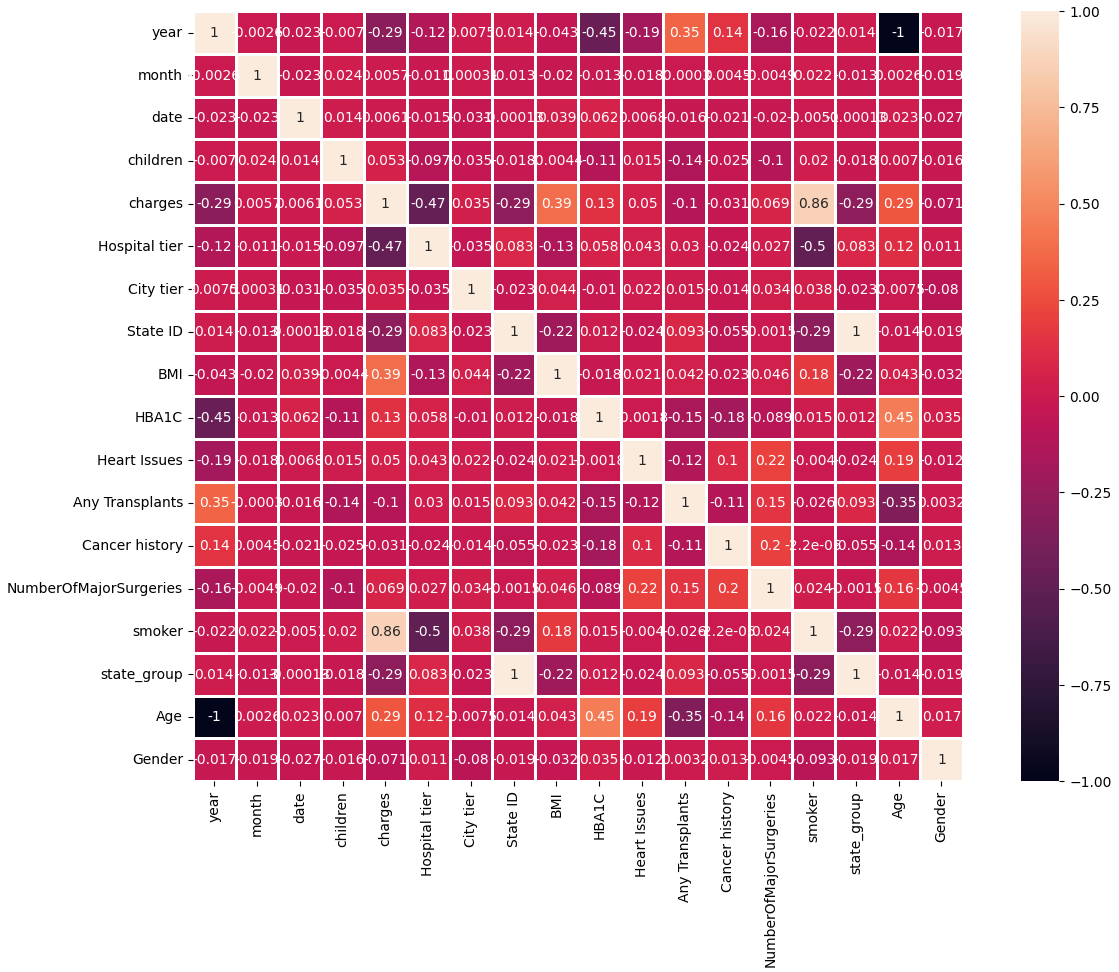
```
2333                    1        0             3    32       0
2334                    1        0             3    32       0

[1755 rows x 18 columns]
```

[ ]:
```
# correlation between predictors to identify highly correlated predictors using␣
↪Heatmap.
```

[ ]:
```
month_dict = {'Jan': 1, 'Feb': 2, 'Mar': 3, 'Apr': 4, 'May': 5, 'Jun': 6, 'Jul':
↪ 7, 'Aug': 8, 'Sep': 9, 'Oct': 10, 'Nov': 11, 'Dec': 12}
df['month'] = df['month'].map(month_dict)
```

[ ]:
```
plt.figure(figsize=(15,10))
sns.heatmap(df.corr(),square=True,annot=True,linewidths=1)
```

[ ]: <Axes: >



[ ]:
```
# MACHINE LEARNING :
```

[ ]:
```
# Regression Model.
```

[ ]:
```
from sklearn.model_selection import train_test_split
```

[ ]:
```
x = df.drop(["charges"], axis=1)
y = df['charges']
x_train, x_test, y_train, y_test = train_test_split(x,y, test_size=.
 ↪20,random_state=10)
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
x_train = sc.fit_transform(x_train)
x_test = sc.fit_transform(x_test)
from sklearn.linear_model import SGDRegressor
```

[ ]:
```
from sklearn.model_selection import GridSearchCV
params = {'alpha': [0.0001, 0.001, 0.01, 0.05, 0.1, 0.2,0.3,0.4,0.5,
0.6,0.7,0.8,0.9,1.0,2.0,3.0,4.0,5.0,6.0,7.0,8.0,
9.0,10.0,20,50,100,500,1000],
'penalty': ['l2', 'l1', 'elasticnet']}
sgd = SGDRegressor()
# Cross Validation
folds = 5
model_cv = GridSearchCV(estimator = sgd,
param_grid = params,
scoring = 'neg_mean_absolute_error',
cv = folds,
return_train_score = True,
verbose = 1)
model_cv.fit(x_train,y_train)
```

```
Fitting 5 folds for each of 84 candidates, totalling 420 fits
```

[ ]:
```
GridSearchCV(cv=5, estimator=SGDRegressor(),
             param_grid={'alpha': [0.0001, 0.001, 0.01, 0.05, 0.1, 0.2, 0.3,
                                   0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 2.0, 3.0,
                                   4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0, 20, 50,
                                   100, 500, 1000],
                         'penalty': ['l2', 'l1', 'elasticnet']},
             return_train_score=True, scoring='neg_mean_absolute_error',
             verbose=1)
```

[ ]:
```
model_cv.best_params_
```

[ ]:
```
{'alpha': 100, 'penalty': 'l1'}
```

[ ]:
```
sgd = SGDRegressor(alpha= 100, penalty= 'l1')
```

[ ]:
```
sgd.fit(x_train, y_train)
```

[ ]:
```
SGDRegressor(alpha=100, penalty='l1')
```

[ ]:
```
sgd.score(x_test, y_test)
```

[ ]:
```
0.8834193279397483
```

[ ]:
```
y_pred = sgd.predict(x_test)
```

[ ]:
```
from sklearn.metrics import mean_squared_error, mean_absolute_error
sgd_mae = mean_absolute_error(y_test, y_pred)
sgd_mse = mean_squared_error(y_test, y_pred)
sgd_rmse = sgd_mse*(1/2.0)
```

[ ]:
```
 print("MAE:", sgd_mae)
print("MSE:", sgd_mse)
print("RMSE:", sgd_rmse)
```

```
MAE: 2803.6752966121044
MSE: 18823121.259697914
RMSE: 9411560.629848957
```

[ ]:
```
importance = sgd.coef_
pd.DataFrame(importance, index = x.columns, columns=['Feature_imp'])
```

[ ]:

	Feature_imp
year	-1781.166092
month	-80.385998
date	-45.058187
children	353.246285
Hospital tier	-1217.079141
City tier	0.000000
State ID	-55.376234
BMI	2734.018350
HBA1C	156.008362
Heart Issues	0.000000
Any Transplants	52.446010
Cancer history	0.000000
NumberOfMajorSurgeries	23.963606
smoker	9561.231409
state_group	-55.376234
Age	1781.166092
Gender	0.000000

[ ]:
```
from sklearn.ensemble import RandomForestRegressor
rf = RandomForestRegressor(n_estimators = 1000, random_state = 42)
rf.fit(x_train, y_train)
```

```
RandomForestRegressor(n_estimators=1000, random_state=42)
score = rf.score(x_test,y_test)
score
```

[ ]: 0.9254277982049066

[ ]:
```
y_pred = rf.predict(x_test)
rf_mae = mean_absolute_error(y_test, y_pred)
rf_mae
```

[ ]: 2000.671016894581

[ ]:
```
from sklearn.ensemble import GradientBoostingRegressor
gbr = GradientBoostingRegressor(n_estimators = 1000, random_state = 42)
gbr.fit(x_train, y_train)
```

[ ]: GradientBoostingRegressor(n_estimators=1000, random_state=42)

[ ]:
```
score = gbr.score(x_test,y_test)
score
```

[ ]: 0.9041039008686199

[ ]:
```
y_pred = gbr.predict(x_test)
gbr_mae = mean_absolute_error(y_test, y_pred)
gbr_mae
```

[ ]: 2567.9912072119014

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

[ ]:
```
Index(['year', 'month', 'date', 'children', 'charges', 'Hospital tier',
       'City tier', 'State ID', 'BMI', 'HBA1C', 'Heart Issues',
       'Any Transplants', 'Cancer history', 'NumberOfMajorSurgeries', 'smoker',
       'state_group', 'Age', 'Gender'],
      dtype='object')
```

[ ]:
```
# Given Case Scenario.
```

[ ]:
```
df1= pd.DataFrame({ 'year' : [1998], 'month' : [12] ,'date': [28],
                    'city_tier' : [1], 'children' :[ 2],
                     'HbA1c' : [5.8],
                     'smoker_yes' : [1],
                     'heart_issues_yes' : [0],
                     'any_transplants_yes' : [0],
                     'numberofmajorsurgeries' :[ 0],
                     'cancer_history_yes' : [1],
                     'hospital_tier' : [1],
```

```
                        'bmi' : [85/(1.70 **2)], 'age' : [25],'Gender' :␣
↪[0],'state_group' : [1],
                        'state_id_R1011' : [1]
                       })
```

[ ]:
```
# Predicting Hospital cost.
```

[ ]:
```
Hospital_cost = []
```

[ ]:
```
Cost1 = sgd.predict(df1)
Hospital_cost.append(Cost1)
Cost2 = rf.predict(df1)
Hospital_cost.append(Cost2)
Cost3 = gbr.predict(df1)
Hospital_cost.append(Cost3)
avg_cost = np.mean(Hospital_cost)
avg_cost
```

[ ]:
```
-1076629.4613465841
```

[ ]:
```

```