Lecture 9

May 25, 2023

1 Data preprocessing

1.1 Import the necessary libraries

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

1.2 Import dataset

[2]:
```
data = pd.read_csv('insurance.csv')
```

1.3 Basic operations on dataset

[3]:
```
data.head()
```

[3]:

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

[4]:
```
data.tail()
```

[4]:

	age	sex	bmi	children	smoker	region	charges
1333	50	male	30.97	3	no	northwest	10600.5483
1334	18	female	31.92	0	no	northeast	2205.9808
1335	18	female	36.85	0	no	southeast	1629.8335
1336	21	female	25.80	0	no	southwest	2007.9450
1337	61	female	29.07	0	yes	northwest	29141.3603

[5]:
```
data.shape
```

[5]:
```
(1338, 7)
```

[6]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   object
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   object
 5   region    1338 non-null   object
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

[7]:
```
data.describe()
```

[7]:

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

[8]:
```
data.mean()
```

```
/var/folders/03/k1p5_v6d69bg7b999gdktlgw0000gn/T/ipykernel_2680/531903386.py:1:
FutureWarning: The default value of numeric_only in DataFrame.mean is
deprecated. In a future version, it will default to False. In addition,
specifying 'numeric_only=None' is deprecated. Select only valid columns or
specify the value of numeric_only to silence this warning.
  data.mean()
```

[8]:
```
age            39.207025
bmi            30.663397
children        1.094918
charges     13270.422265
dtype: float64
```

[9]:
```
data.median()
```

```
/var/folders/03/k1p5_v6d69bg7b999gdktlgw0000gn/T/ipykernel_2680/4184645713.py:1:
FutureWarning: The default value of numeric_only in DataFrame.median is
deprecated. In a future version, it will default to False. In addition,
specifying 'numeric_only=None' is deprecated. Select only valid columns or
specify the value of numeric_only to silence this warning.
```

```
  data.median()
```

[9]:
```
age           39.000
bmi           30.400
children       1.000
charges     9382.033
dtype: float64
```

[10]:
```
data.mode()
```

[10]:

	age	sex	bmi	children	smoker	region	charges
0	18	male	32.3	0	no	southeast	1639.5631

[11]:
```
data.var()
```

```
/var/folders/03/k1p5_v6d69bg7b999gdktlgw0000gn/T/ipykernel_2680/445316826.py:1:
FutureWarning: The default value of numeric_only in DataFrame.var is deprecated.
In a future version, it will default to False. In addition, specifying
'numeric_only=None' is deprecated. Select only valid columns or specify the
value of numeric_only to silence this warning.
  data.var()
```

[11]:
```
age         1.974014e+02
bmi         3.718788e+01
children    1.453213e+00
charges     1.466524e+08
dtype: float64
```

[12]:
```
data.std()
```

```
/var/folders/03/k1p5_v6d69bg7b999gdktlgw0000gn/T/ipykernel_2680/2723740006.py:1:
FutureWarning: The default value of numeric_only in DataFrame.std is deprecated.
In a future version, it will default to False. In addition, specifying
'numeric_only=None' is deprecated. Select only valid columns or specify the
value of numeric_only to silence this warning.
  data.std()
```

[12]:
```
age           14.049960
bmi            6.098187
children       1.205493
charges    12110.011237
dtype: float64
```

[13]:
```
data.isnull().sum()
```

[13]:
```
age         0
sex         0
bmi         0
children    0
```

```
smoker      0
region      0
charges     0
dtype: int64
```

[14]:
```
data.sex.value_counts()
```

[14]:
```
male      676
female    662
Name: sex, dtype: int64
```

1.4 Handling missing value

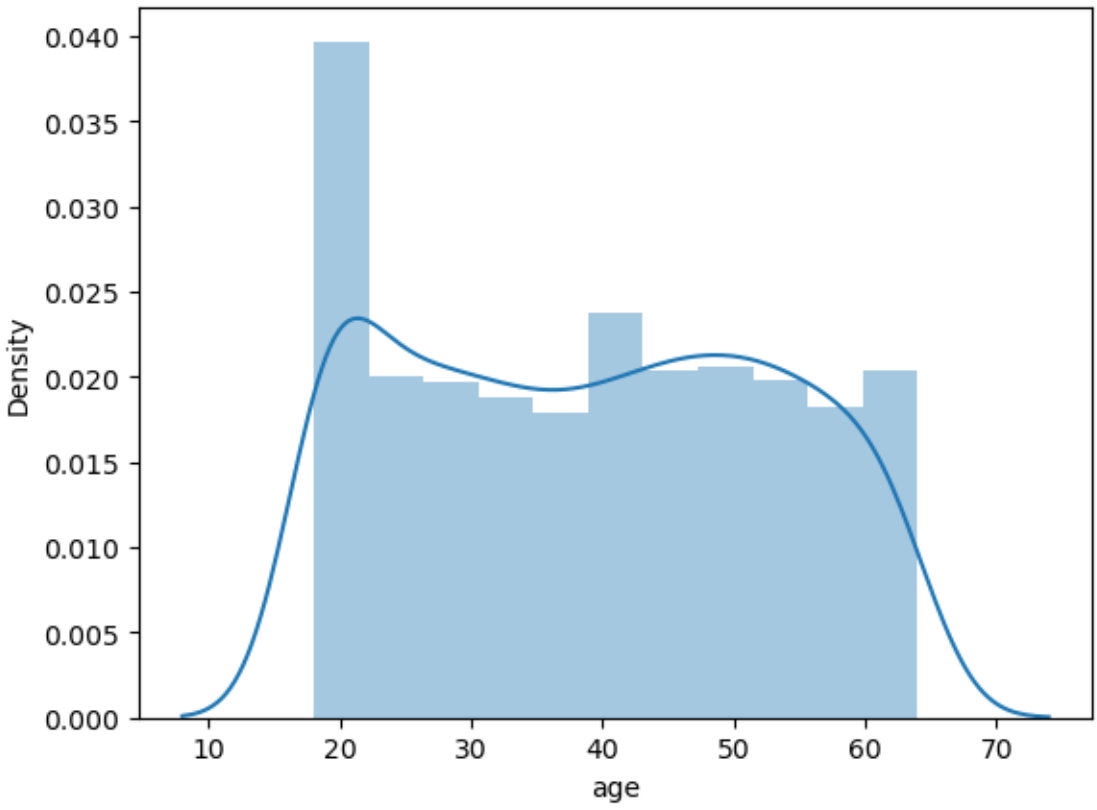
[15]:
```
data['age'].fillna(data['age'].mean(), inplace=True)
```

1.5 Data visualization

[16]:
```
sns.distplot(data.age)
```

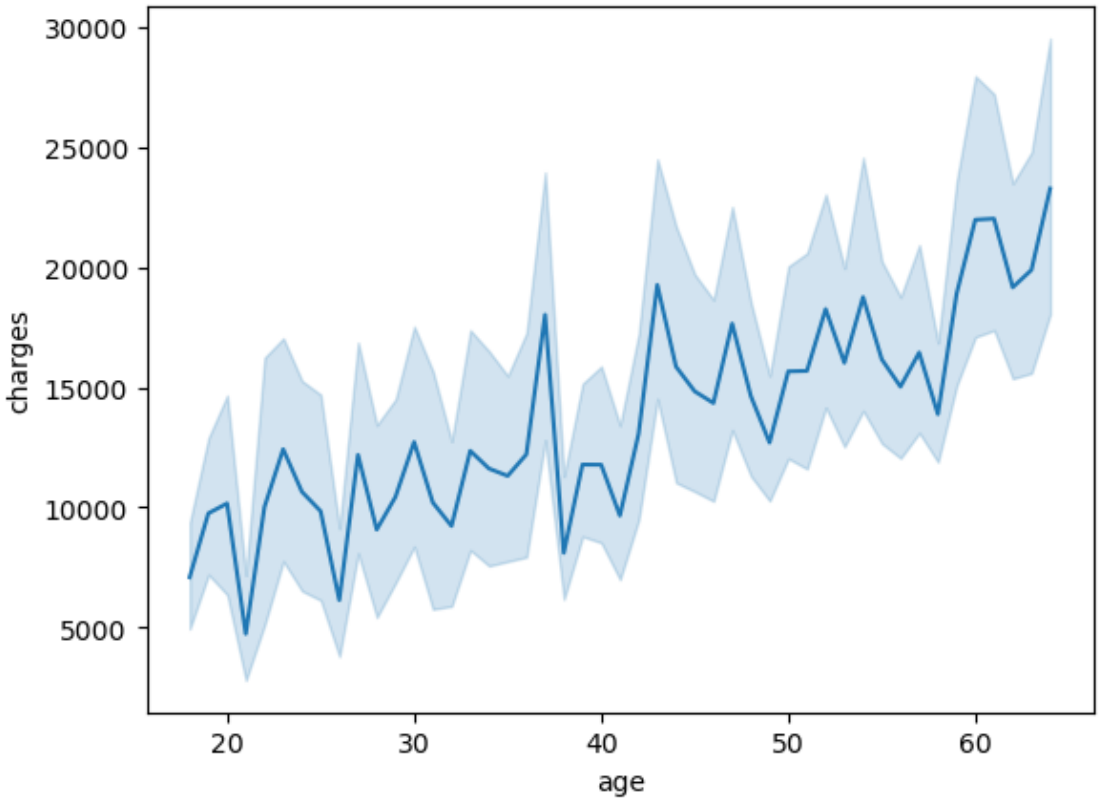
```
/var/folders/03/k1p5_v6d69bg7b999gdktlgw0000gn/T/ipykernel_2680/4156840497.py:1:
UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(data.age)
```

[16]:
```
<Axes: xlabel='age', ylabel='Density'>
```



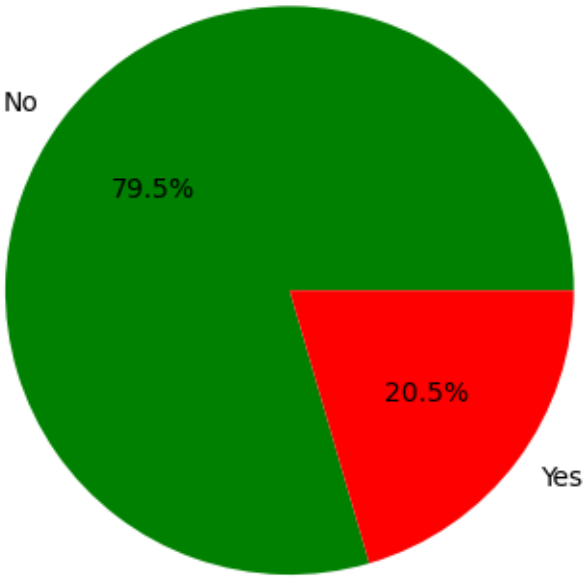
[17]:
```
sns.lineplot(x=data.age, y=data.charges)
```

[17]:
```
<Axes: xlabel='age', ylabel='charges'>
```



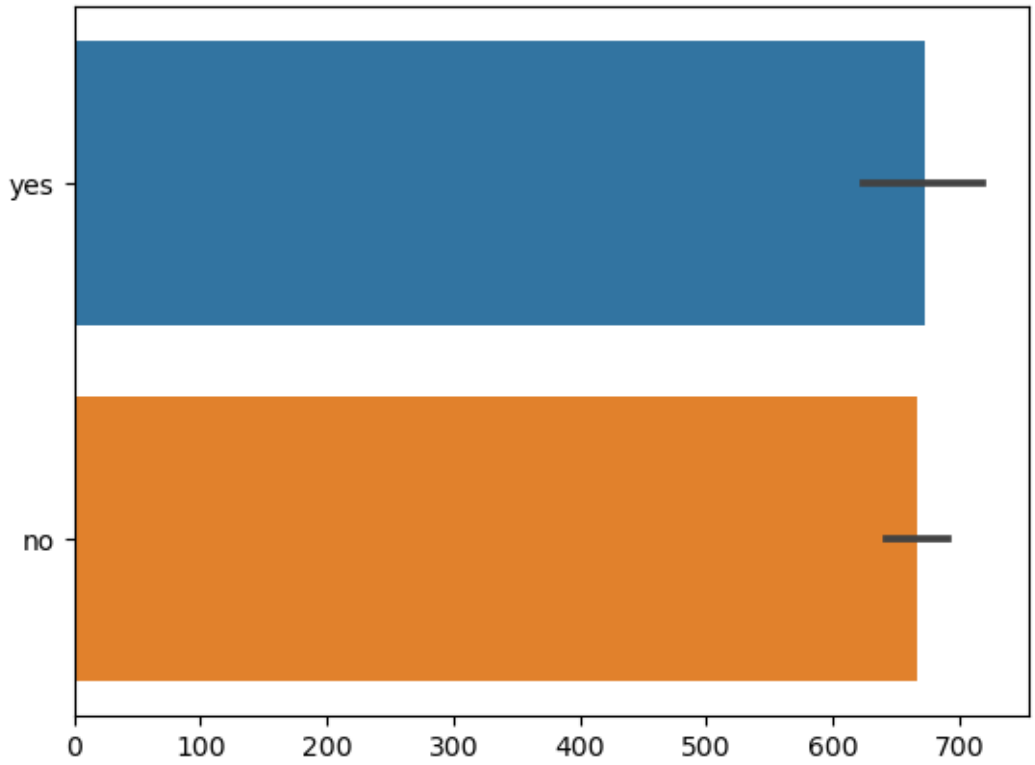
[18]:
```
plt.pie(data.smoker.value_counts(), colors=['Green', 'Red'], labels = ['No',␣
↪'Yes'], autopct='%.1f%%')
```

[18]:
```
([<matplotlib.patches.Wedge at 0x14245f640>,
  <matplotlib.patches.Wedge at 0x14245f580>],
 [Text(-0.8801026332278259, 0.6598631335250117, 'No'),
  Text(0.8801026332278258, -0.6598631335250118, 'Yes')],
 [Text(-0.48005598176063224, 0.3599253455590972, '79.5%'),
  Text(0.48005598176063213, -0.3599253455590973, '20.5%')])
```



[19]:
```
sns.barplot(x=data.smoker.index, y=data.smoker.values)
```

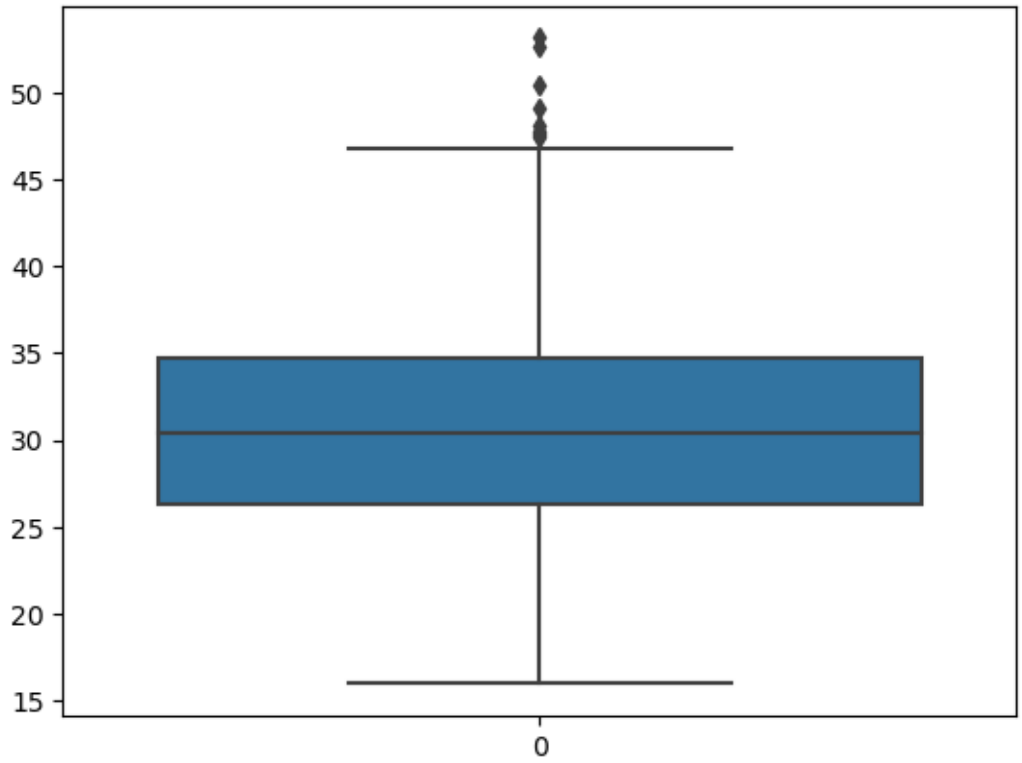
[19]: <Axes: >



1.6 Removing outliers

[20]:
```
sns.boxplot(data.bmi)
```

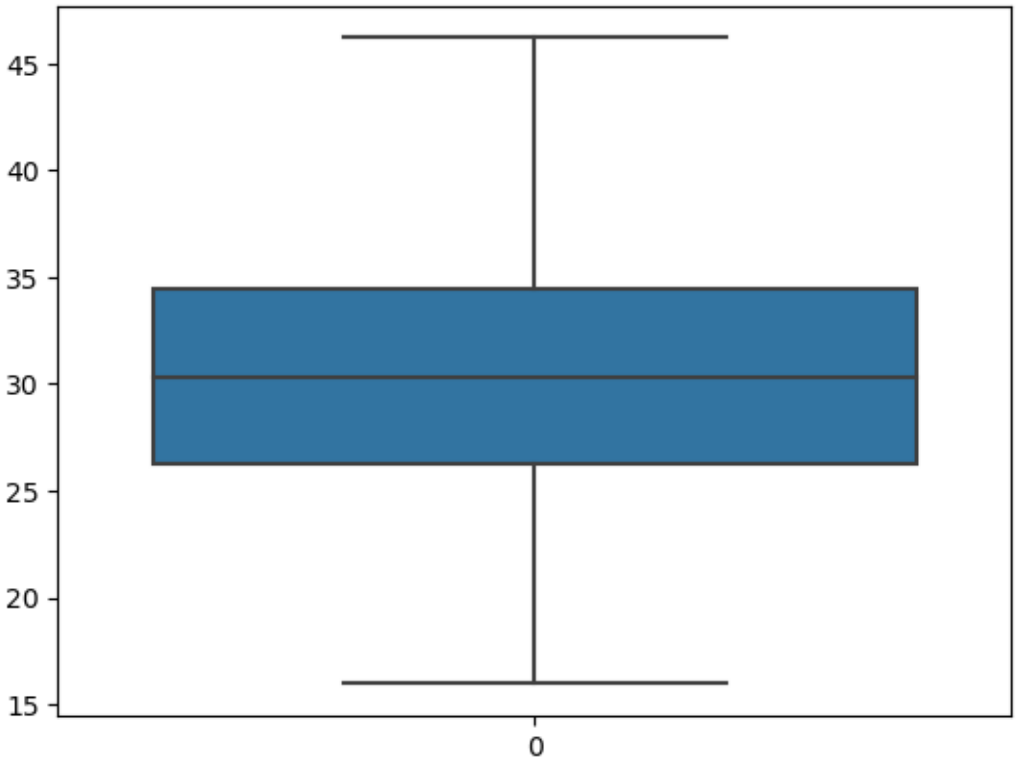
[20]: <Axes: >



[21]:
```
quant99 = data.bmi.quantile(0.99)
data = data[data.bmi<quant99]
```

[22]:
```
sns.boxplot(data.bmi)
```

[22]: <Axes: >



1.7 Encoding techniques

1.7.1 Label encoding

[23]:
```
from sklearn.preprocessing import LabelEncoder
```

[24]:
```
le = LabelEncoder()
```

[25]:
```
data.sex = le.fit_transform(data.sex)
```

[26]:
```
data.head()
```

[26]:

	age	sex	bmi	children	smoker	region	charges
0	19	0	27.900	0	yes	southwest	16884.92400
1	18	1	33.770	1	no	southeast	1725.55230
2	28	1	33.000	3	no	southeast	4449.46200
3	33	1	22.705	0	no	northwest	21984.47061
4	32	1	28.880	0	no	northwest	3866.85520

[27]:
```
data.smoker = le.fit_transform(data.smoker)
```

[28]:
```
data.head()
```

[28]:

	age	sex	bmi	children	smoker	region	charges
0	19	0	27.900	0	1	southwest	16884.92400
1	18	1	33.770	1	0	southeast	1725.55230
2	28	1	33.000	3	0	southeast	4449.46200
3	33	1	22.705	0	0	northwest	21984.47061
4	32	1	28.880	0	0	northwest	3866.85520

1.7.2 One Hot Encoding

[29]:
```
data_main = pd.get_dummies(data, columns=['region'])
```

[30]:
```
data_main
```

[30]:

	age	sex	bmi	children	smoker	charges	region_northeast	\
0	19	0	27.900	0	1	16884.92400	0	
1	18	1	33.770	1	0	1725.55230	0	
2	28	1	33.000	3	0	4449.46200	0	
3	33	1	22.705	0	0	21984.47061	0	
4	32	1	28.880	0	0	3866.85520	0	
…	…	…	…	…	…	…	…	
1333	50	1	30.970	3	0	10600.54830	0	
1334	18	0	31.920	0	0	2205.98080	1	
1335	18	0	36.850	0	0	1629.83350	0	
1336	21	0	25.800	0	0	2007.94500	0	
1337	61	0	29.070	0	1	29141.36030	0	

	region_northwest	region_southeast	region_southwest
0	0	0	1
1	0	1	0
2	0	1	0
3	1	0	0
4	1	0	0
…	…	…	…
1333	1	0	0
1334	0	0	0
1335	0	1	0
1336	0	0	1
1337	1	0	0

[1324 rows x 10 columns]

1.8 Correlation

[31]:
```
data_main.corr()
```

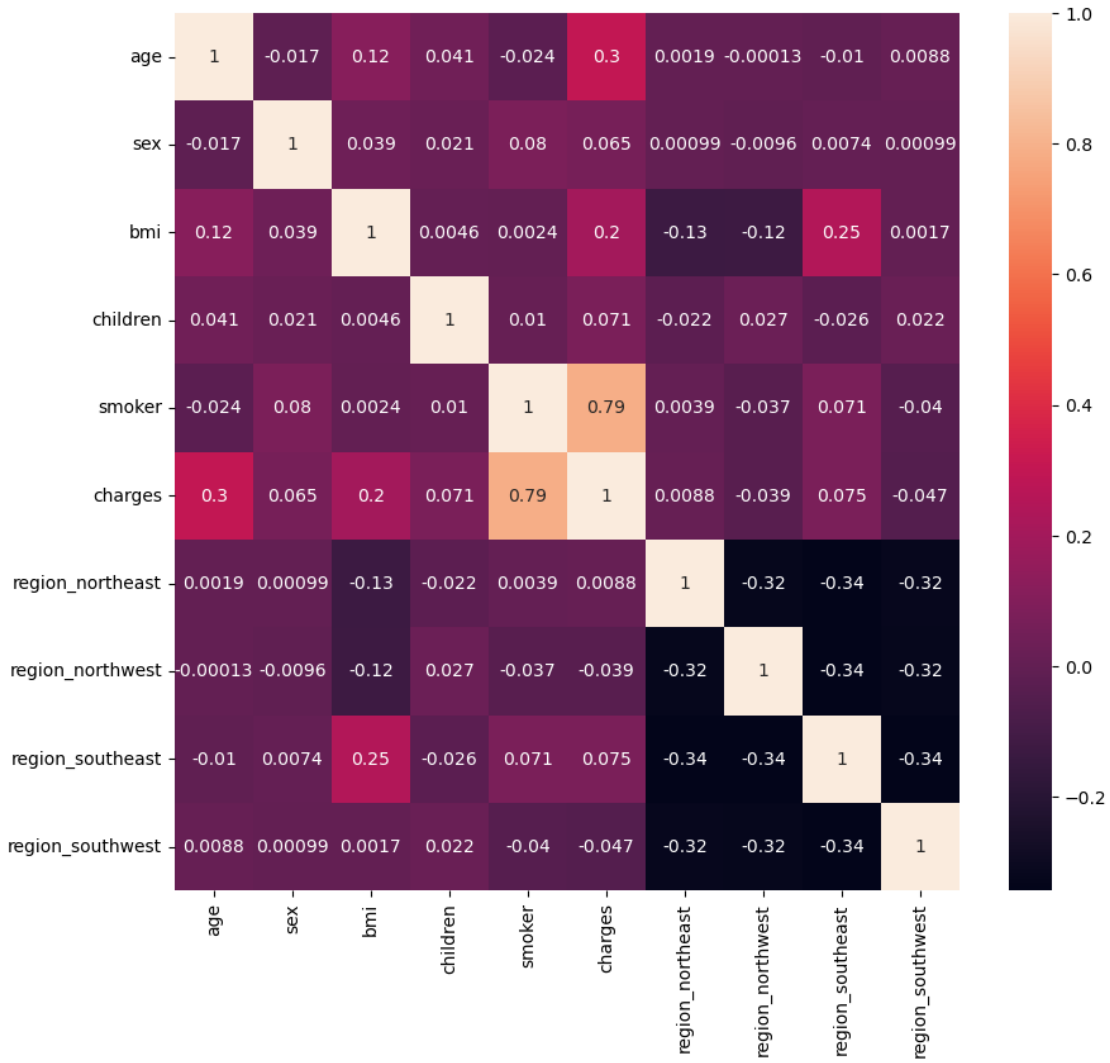
[31]:

	age	sex	bmi	children	smoker	charges	\
age	1.000000	-0.016501	0.115670	0.040913	-0.024410	0.301754	

```
sex              -0.016501  1.000000  0.038665  0.020772  0.080410  0.065004
bmi               0.115670  0.038665  1.000000  0.004627  0.002354  0.200882
children          0.040913  0.020772  0.004627  1.000000  0.010016  0.071393
smoker           -0.024410  0.080410  0.002354  0.010016  1.000000  0.785872
charges           0.301754  0.065004  0.200882  0.071393  0.785872  1.000000
region_northeast  0.001921  0.000986 -0.132316 -0.022324  0.003868  0.008765
region_northwest -0.000131 -0.009573 -0.124586  0.027045 -0.037068 -0.038750
region_southeast -0.010289  0.007403  0.248115 -0.025542  0.070897  0.074616
region_southwest  0.008804  0.000986  0.001697  0.021521 -0.039720 -0.046761

                  region_northeast  region_northwest  region_southeast  \
age                       0.001921         -0.000131         -0.010289
sex                       0.000986         -0.009573          0.007403
bmi                      -0.132316         -0.124586          0.248115
children                 -0.022324          0.027045         -0.025542
smoker                    0.003868         -0.037068          0.070897
charges                   0.008765         -0.038750          0.074616
region_northeast          1.000000         -0.323999         -0.342501
region_northwest         -0.323999          1.000000         -0.343904
region_southeast         -0.342501         -0.343904          1.000000
region_southwest         -0.322677         -0.323999         -0.342501

                  region_southwest
age                       0.008804
sex                       0.000986
bmi                       0.001697
children                  0.021521
smoker                   -0.039720
charges                  -0.046761
region_northeast         -0.322677
region_northwest         -0.323999
region_southeast         -0.342501
region_southwest          1.000000
```

[32]:
```
plt.figure(figsize=(10,9))
sns.heatmap(data_main.corr(), annot=True)
```

[32]: <Axes: >



[33]:
```
data_main.head()
```

[33]:

	age	sex	bmi	children	smoker	charges	region_northeast	region_northwest	region_southeast	region_southwest
0	19	0	27.900	0	1	16884.92400	0	0	0	1
1	18	1	33.770	1	0	1725.55230	0	0	1	0
2	28	1	33.000	3	0	4449.46200	0	0	1	0
3	33	1	22.705	0	0	21984.47061	0	1	0	0
4	32	1	28.880	0	0	3866.85520	0	0	0	0

```
4                 1                 0                 0
```

1.9 X and Y Split

[34]:
```
y = data_main['charges']
```

[35]:
```
y.head()
```

[35]:
```
0    16884.92400
1     1725.55230
2     4449.46200
3    21984.47061
4     3866.85520
Name: charges, dtype: float64
```

[36]:
```
X = data_main.drop(columns=['charges'], axis=1)
```

[37]:
```
X.head()
```

[37]:
```
   age  sex     bmi  children  smoker  region_northeast  region_northwest  \
0   19    0  27.900         0       1                 0                 0
1   18    1  33.770         1       0                 0                 0
2   28    1  33.000         3       0                 0                 0
3   33    1  22.705         0       0                 0                 1
4   32    1  28.880         0       0                 0                 1

   region_southeast  region_southwest
0                 0                 1
1                 1                 0
2                 1                 0
3                 0                 0
4                 0                 0
```

1.10 Scaling

StandardScaler –> mean=0 std=1 MinMaxScaler –> scale between 0 to 1

[38]:
```
from sklearn.preprocessing import MinMaxScaler
scale = MinMaxScaler()
```

[39]:
```
name = X.columns
X_scaled = scale.fit_transform(X)
```

[40]:
```
X_scaled
```

[40]:
```
array([[0.02173913, 0.        , 0.39484127, …, 0.        , 0.        ,
        1.        ],
       [0.        , 1.        , 0.58895503, …, 0.        , 1.        ,
```

```
        0.        ],
       [0.2173913 , 1.        , 0.56349206, …, 0.        , 1.        ,
        0.        ],
       …,
       [0.        , 0.        , 0.69080688, …, 0.        , 1.        ,
        0.        ],
       [0.06521739, 0.        , 0.32539683, …, 0.        , 0.        ,
        1.        ],
       [0.93478261, 0.        , 0.43353175, …, 1.        , 0.        ,
        0.        ]])
```

[41]:
```
X = pd.DataFrame(X_scaled, columns=name)
X
```

[41]:

	age	sex	bmi	children	smoker	region_northeast	region_northwest	region_southeast	region_southwest
0	0.021739	0.0	0.394841	0.0	1.0	0.0	0.0	0.0	1.0
1	0.000000	1.0	0.588955	0.2	0.0	0.0	0.0	1.0	0.0
2	0.217391	1.0	0.563492	0.6	0.0	0.0	0.0	1.0	0.0
3	0.326087	1.0	0.223049	0.0	0.0	0.0	1.0	0.0	0.0
4	0.304348	1.0	0.427249	0.0	0.0	0.0	1.0	0.0	0.0
…	…	…	…	…	…	…	…	…	…
1319	0.695652	1.0	0.496362	0.6	0.0	0.0	1.0	0.0	0.0
1320	0.000000	0.0	0.527778	0.0	0.0	1.0	0.0	0.0	1.0
1321	0.000000	0.0	0.690807	0.0	0.0	0.0	0.0	1.0	0.0
1322	0.065217	0.0	0.325397	0.0	0.0	0.0	0.0	0.0	1.0
1323	0.934783	0.0	0.433532	0.0	1.0	0.0	1.0	0.0	0.0

[1324 rows x 9 columns]

1.11 Train-Test Split

[42]:
```
from sklearn.model_selection import train_test_split
```

[43]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,␣
↪random_state=0)
```

[44]:
```
X_train
```

[44]:
```
           age  sex       bmi  children  smoker  region_northeast  \
49    0.391304  1.0  0.636243       0.2     1.0               0.0
672   0.065217  0.0  0.047619       0.2     0.0               0.0
911   0.956522  0.0  0.570106       0.0     0.0               0.0
936   0.413043  1.0  0.603175       0.2     1.0               1.0
769   0.586957  1.0  0.788525       0.0     0.0               1.0
…          …   …         …         …       …                 …
763   0.760870  0.0  0.355159       0.4     0.0               0.0
835   0.847826  0.0  0.458003       0.0     1.0               0.0
1216  0.869565  1.0  0.474372       0.0     0.0               1.0
559   0.434783  0.0  0.813657       0.2     0.0               0.0
684   0.043478  1.0  0.543485       0.2     0.0               0.0

      region_northwest  region_southeast  region_southwest
49                 0.0               1.0               0.0
672                0.0               0.0               1.0
911                0.0               0.0               1.0
936                0.0               0.0               0.0
769                0.0               0.0               0.0
…                  …                 …                 …
763                0.0               0.0               1.0
835                0.0               1.0               0.0
1216               0.0               0.0               0.0
559                1.0               0.0               0.0
684                1.0               0.0               0.0

[1059 rows x 9 columns]
```

[45]:
```
y_train
```

[45]:
```
49      38709.17600
680      2585.26900
921     13462.52000
947     39047.28500
777      7448.40395
            …
771     11150.78000
843     27533.91290
1229    11938.25595
566      6373.55735
692      2362.22905
Name: charges, Length: 1059, dtype: float64
```

[46]:
```
X_test
```

[46]:
```
           age  sex       bmi  children  smoker  region_northeast  \
1294  0.304348  1.0  0.402116       0.8     1.0               0.0
406   0.304348  1.0  0.465278       0.2     0.0               0.0
1062  0.913043  0.0  0.078538       0.0     0.0               1.0
202   0.195652  0.0  0.665344       0.0     1.0               0.0
1191  0.000000  0.0  0.374339       0.6     1.0               0.0
…          …  …         …         …       …                 …
194   0.021739  1.0  0.483796       0.0     0.0               0.0
240   0.326087  0.0  0.204200       0.2     0.0               1.0
1158  0.543478  0.0  0.355159       0.4     1.0               0.0
563   0.282609  0.0  0.434524       0.0     0.0               0.0
1265  0.456522  1.0  0.461806       0.2     1.0               1.0

      region_northwest  region_southeast  region_southwest
1294               1.0               0.0               0.0
406                0.0               1.0               0.0
1062               0.0               0.0               0.0
202                0.0               1.0               0.0
1191               0.0               1.0               0.0
…                  …                 …                 …
194                1.0               0.0               0.0
240                0.0               0.0               0.0
1158               0.0               0.0               1.0
563                0.0               0.0               1.0
1265               0.0               0.0               0.0

[265 rows x 9 columns]
```

[47]:
```
y_test
```

[47]:
```
1307    21472.47880
409      4074.45370
1074    13204.28565
203     37133.89820
1204    18223.45120
           …
195      1639.56310
241      5354.07465
1171    22478.60000
570      3761.29200
1278    22462.04375
Name: charges, Length: 265, dtype: float64
```