```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import warnings

df=pd.read_csv("insurance.csv")

df.head(5)
```

	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

Basic INFO

```
df.shape
```

```
(1338, 7)
```

```
df.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
```

```
df.describe()
```

	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

```
df['region'].unique()
```

```
array(['southwest', 'southeast', 'northwest', 'northeast'],
dtype=object)
```

```
df['children'].unique()
```

```
array([0, 1, 3, 2, 5, 4])
```

```
df['sex'].unique()
```

```
array(['female', 'male'], dtype=object)
```

```
df['smoker'].unique()
```

```
array(['yes', 'no'], dtype=object)
```

```
df.isna().sum()
```

```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

```
df['region_southwest']=df['region'].apply(lambda x : 1 if x
=="southwest" else 0)
```

```
df.head(5)
```

	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	

	region_southwest
0	1
1	0
2	0
3	0
4	0

```
df['region_southeast']=df['region'].apply(lambda x : 1 if x
=="southeast" else 0)
```

```
df['region_northwest']=df['region'].apply(lambda x : 1 if x
=="northwest" else 0)
```

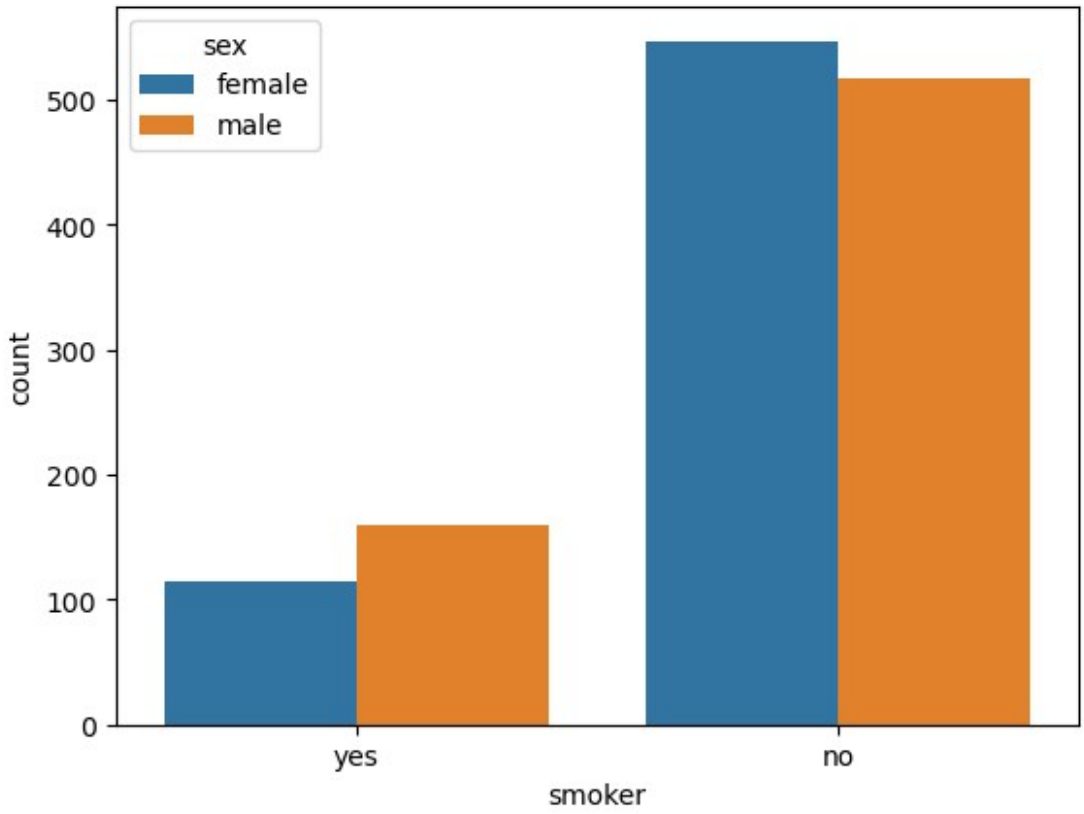
```
df['region_northeast']=df['region'].apply(lambda x : 1 if x
=="northeast" else 0)

df.head(5)
```

```
   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

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

visualisation

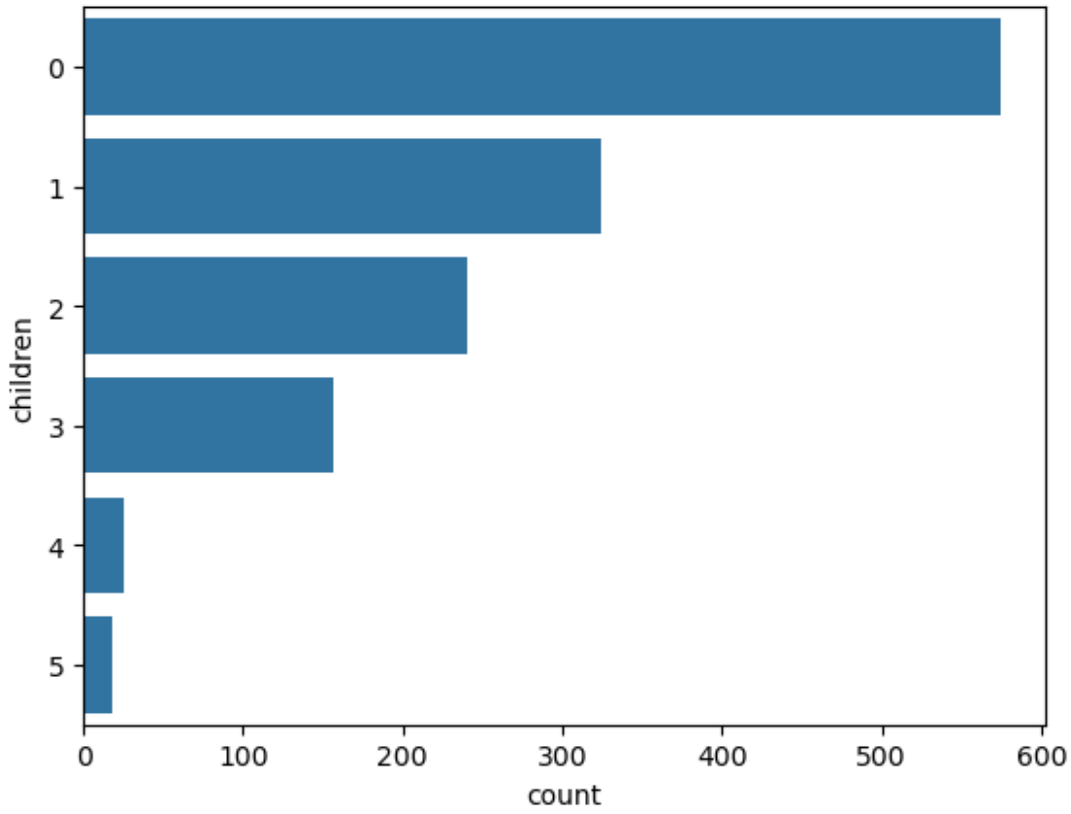
```
sns.countplot(x=df['smoker'],hue=df['sex'])
```

```
<Axes: xlabel='smoker', ylabel='count'>
```



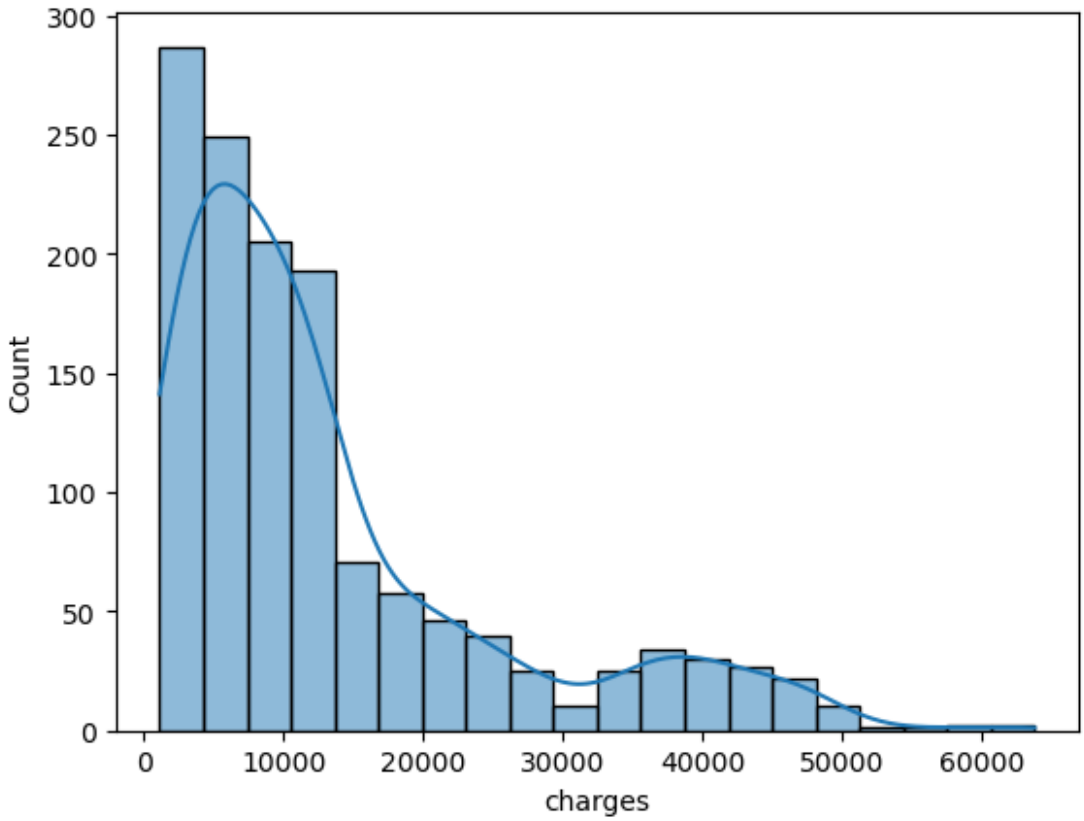

```
sns.countplot(y=df['children'])
```

```
<Axes: xlabel='count', ylabel='children'>
```



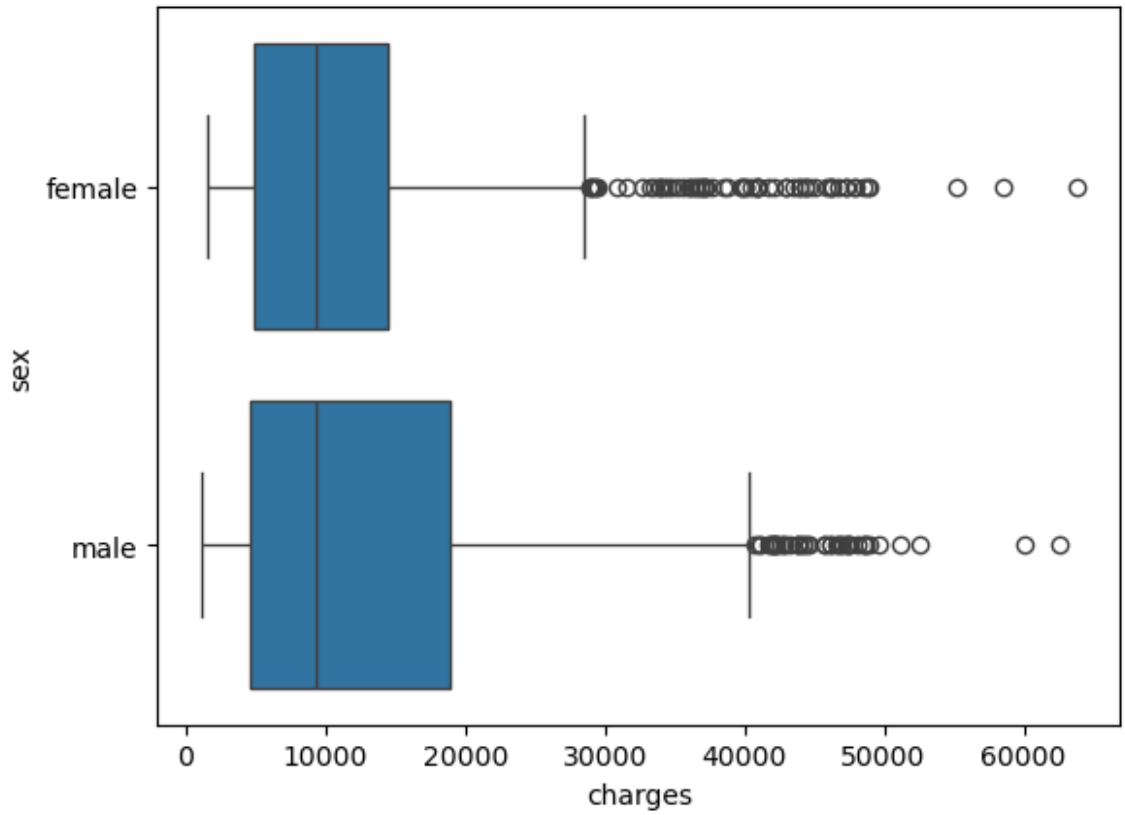

```
sns.histplot(df['charges'],bins=20,kde=True)
```

```
<Axes: xlabel='charges', ylabel='Count'>
```

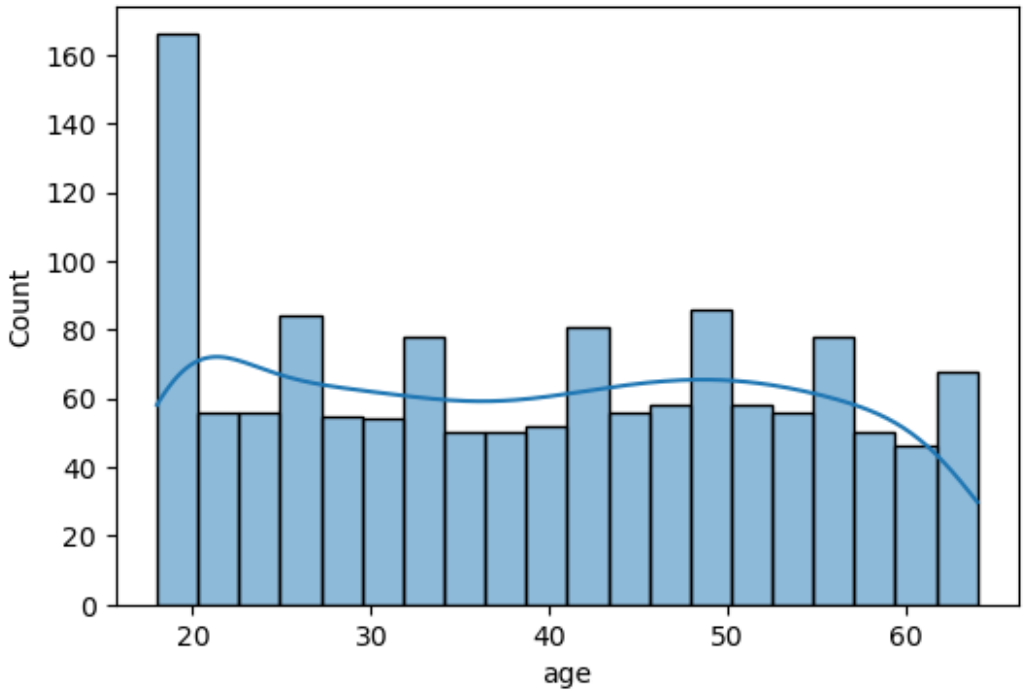


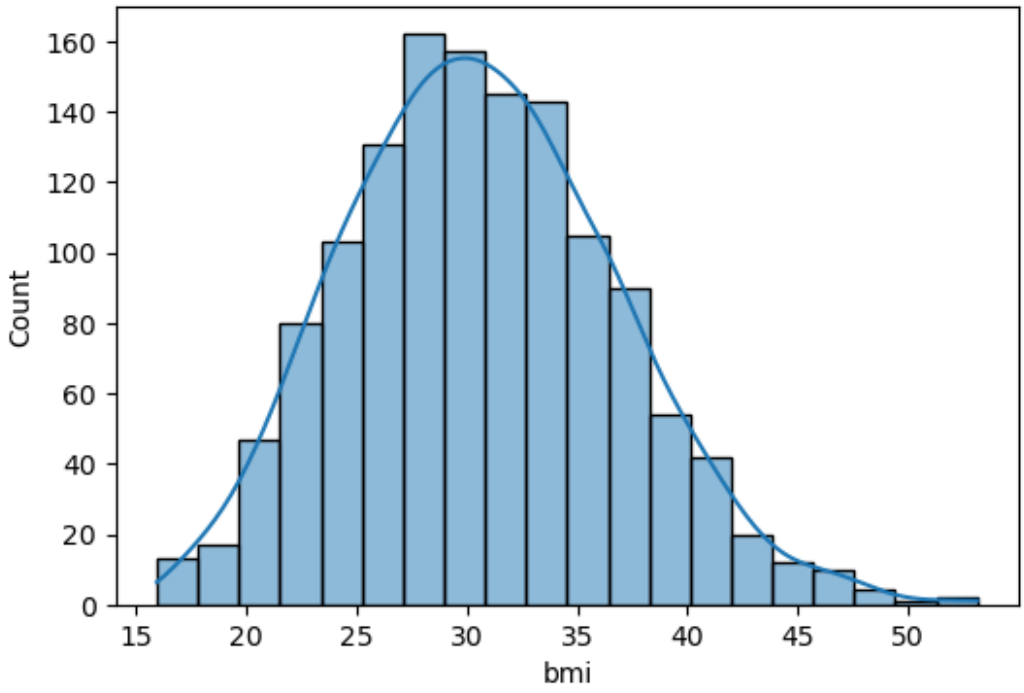

```
sns.boxplot(x='charges' , y='sex' ,data=df)
```

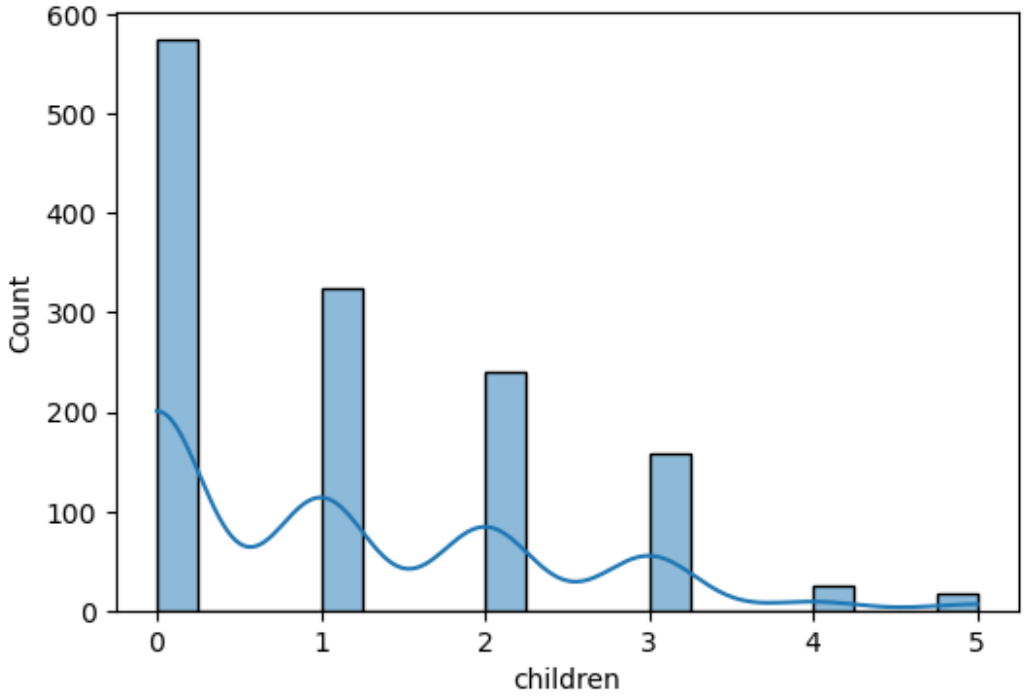
```
<Axes: xlabel='charges', ylabel='sex'>
```

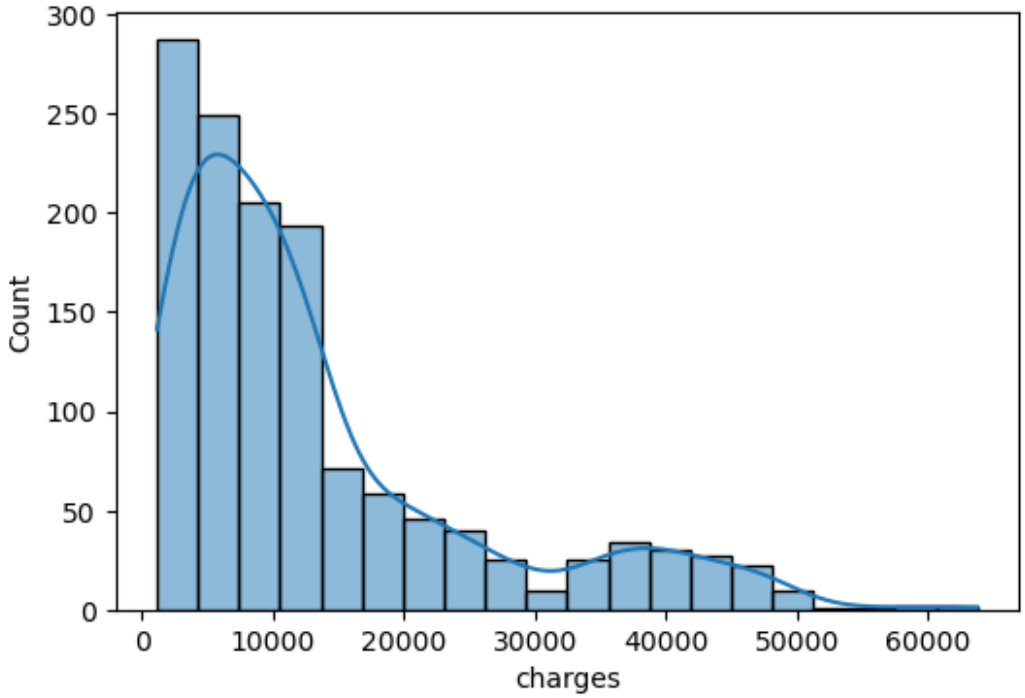



```
numeric_columns=['age','bmi','children','charges']

for col in numeric_columns:
    plt.figure(figsize=(6,4))
    sns.histplot(df[col],kde=True,bins=20)
```

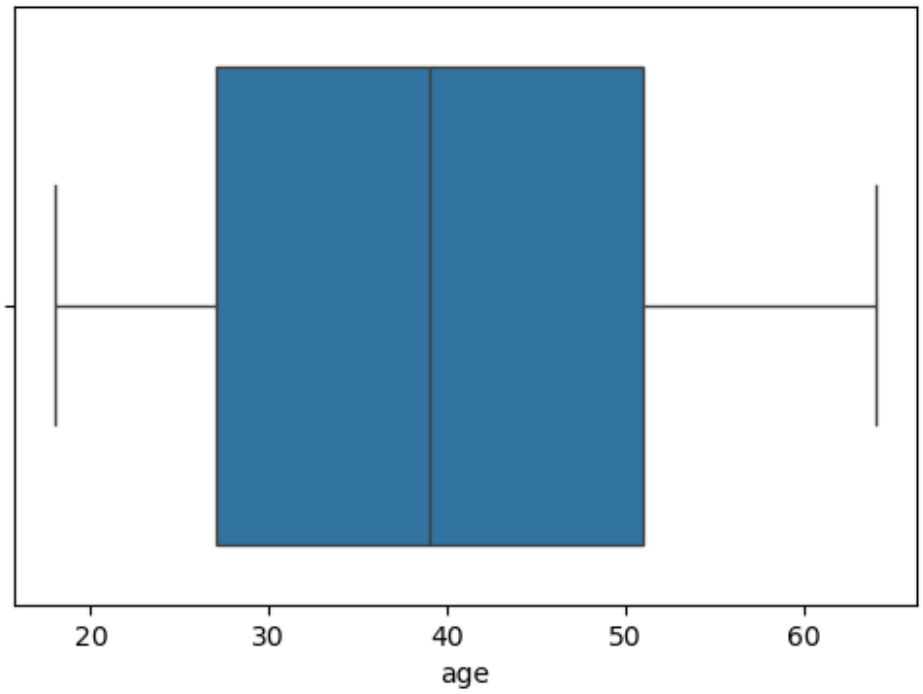

Count
160
140
120
100
80
60
40
20
0
20
30
40
50
60
age

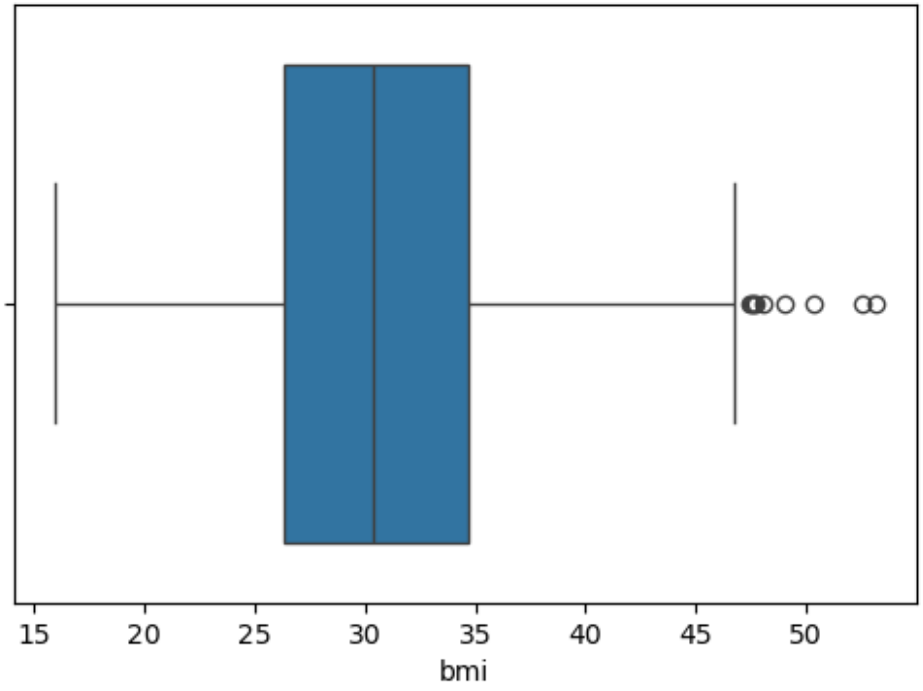

Count
160
140
120
100
80
60
40
20
0
15
20
25
30
35
40
45
50
bmi

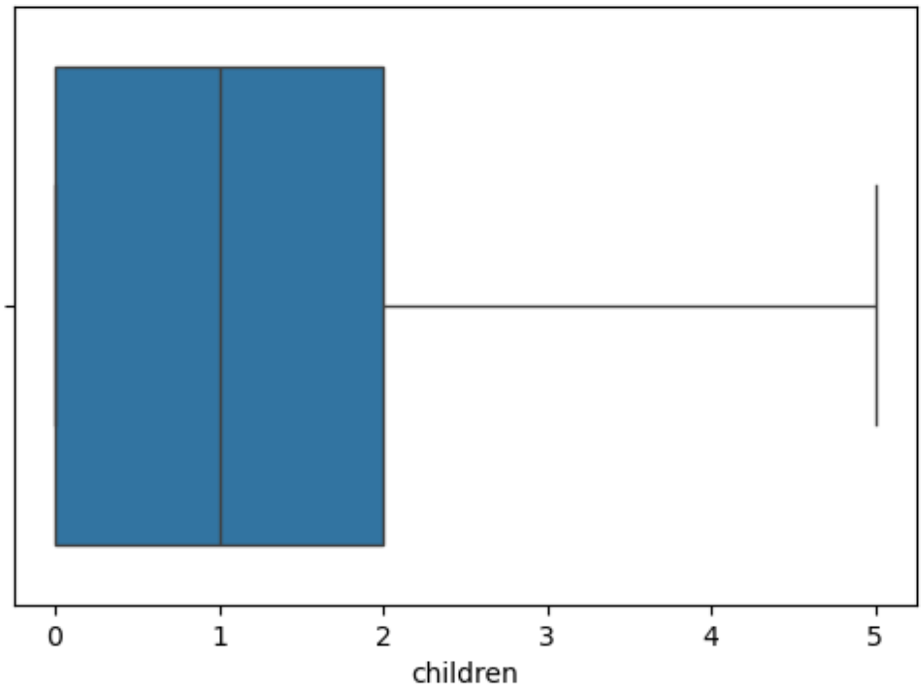


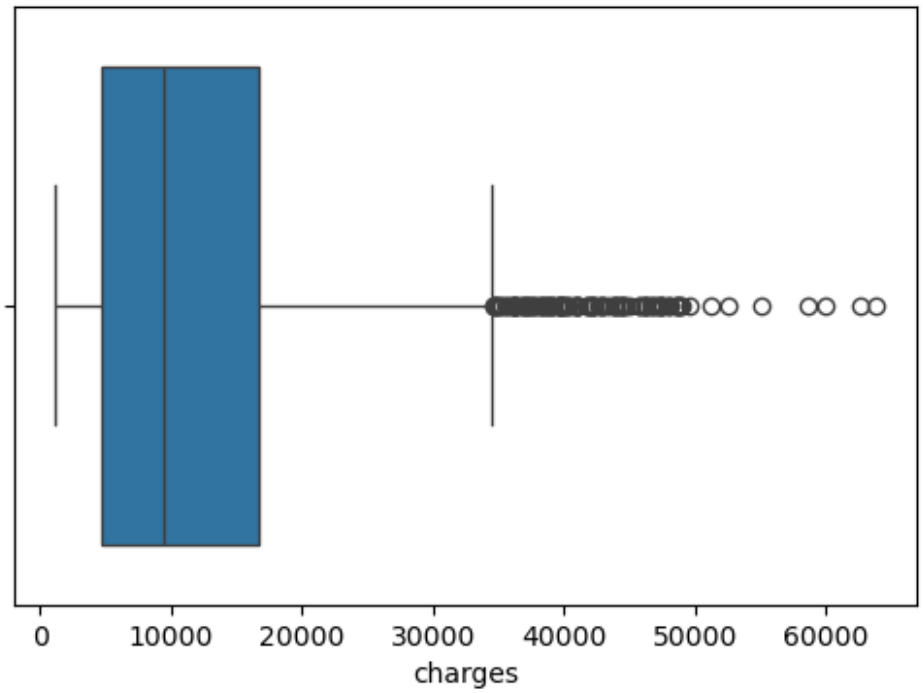



```
for col in numeric_columns:
    plt.figure(figsize=(6,4))
    sns.boxplot(x=df[col])
```


20
30
40
50
60
age

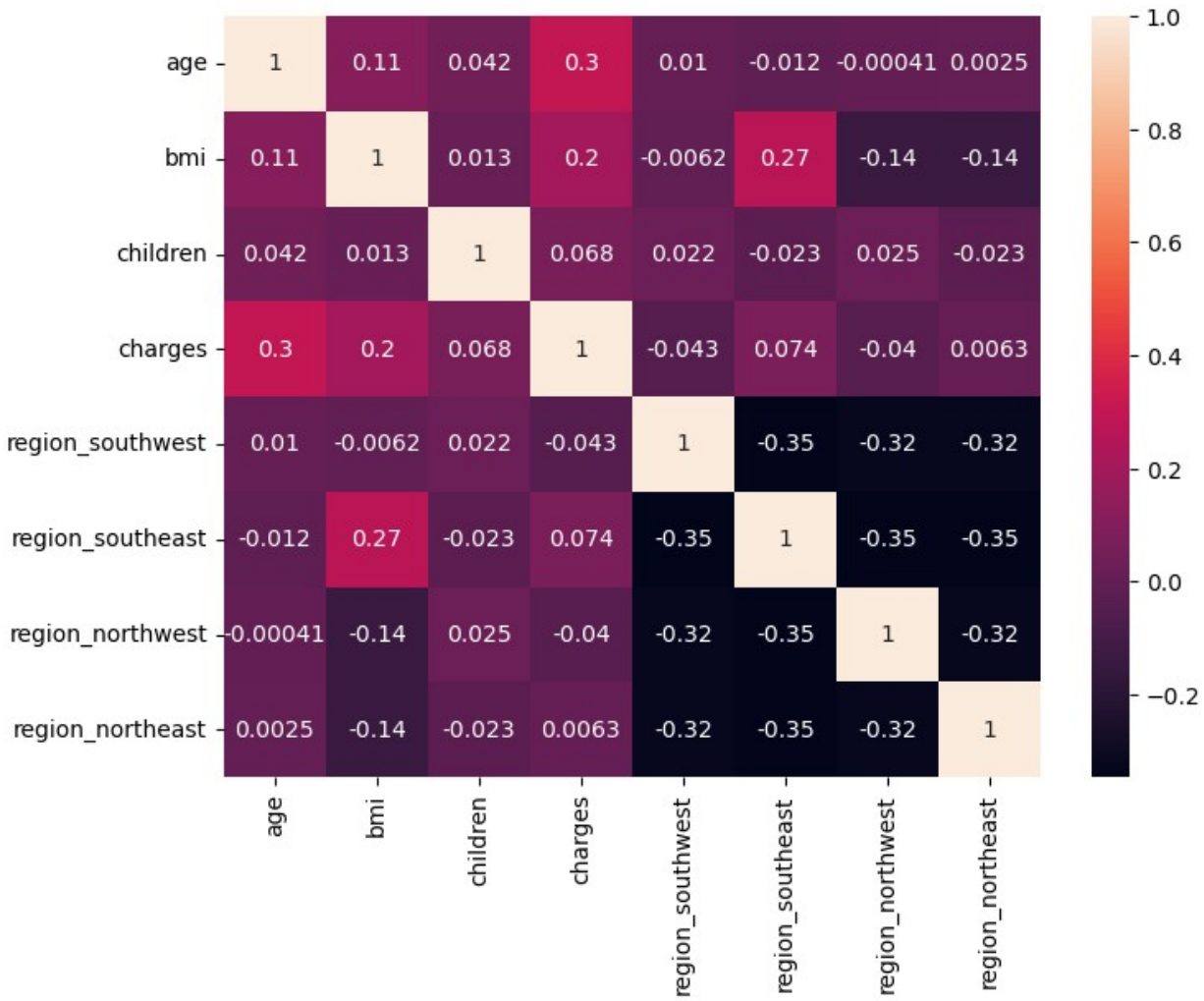

15
20
25
30
35
40
45
50
bmi






```
plt.figure(figsize=(8,6))
sns.heatmap(df.corr(numeric_only=True),annot=True)
```

```
<Axes: >
```



data cleaning and processing

```
df_clean=df.copy()

df_clean.head(5)
```

```
   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

   region_southwest  region_southeast  region_northwest
region_northeast
0                 1                 0                 0
0
1                 0                 1                 0
0
```

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

```
df_clean.drop_duplicates(inplace=True)
```

```
df.shape
```

```
(1338, 11)
```

```
df_clean['smoker']=df_clean['smoker'].apply(lambda x : 1 if x == "yes"
else 0)
```

```
df_clean['sex']=df_clean['sex'].apply(lambda x : 1 if x == "male" else
0)
```

```
df_clean.head(5)
```

```
   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

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

```
df_clean.drop(columns=["region"])
```

```
      age  sex     bmi  children  smoker      charges
region_southwest  \
0      19    0  27.900         0       1  16884.92400
1
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
0
1335   18    0  36.850         0       0   1629.83350
0
1336   21    0  25.800         0       0   2007.94500
1
1337   61    0  29.070         0       1  29141.36030
0

      region_southeast  region_northwest  region_northeast
0                    0                 0                 0
1                    1                 0                 0
2                    1                 0                 0
3                    0                 1                 0
4                    0                 1                 0
...                ...               ...               ...
1333                 0                 1                 0
1334                 0                 0                 1
1335                 1                 0                 0
1336                 0                 0                 0
1337                 0                 1                 0

[1337 rows x 10 columns]
```

```
df_clean.rename(columns={'sex':'is_male','smoker':'is_smoker'},inplace
=True)
# df_clean=df_clean.drop(columns=['region'])

# df_clean.to_csv("INSURANCE_CLEANED.csv",index=False)

df_clean
```

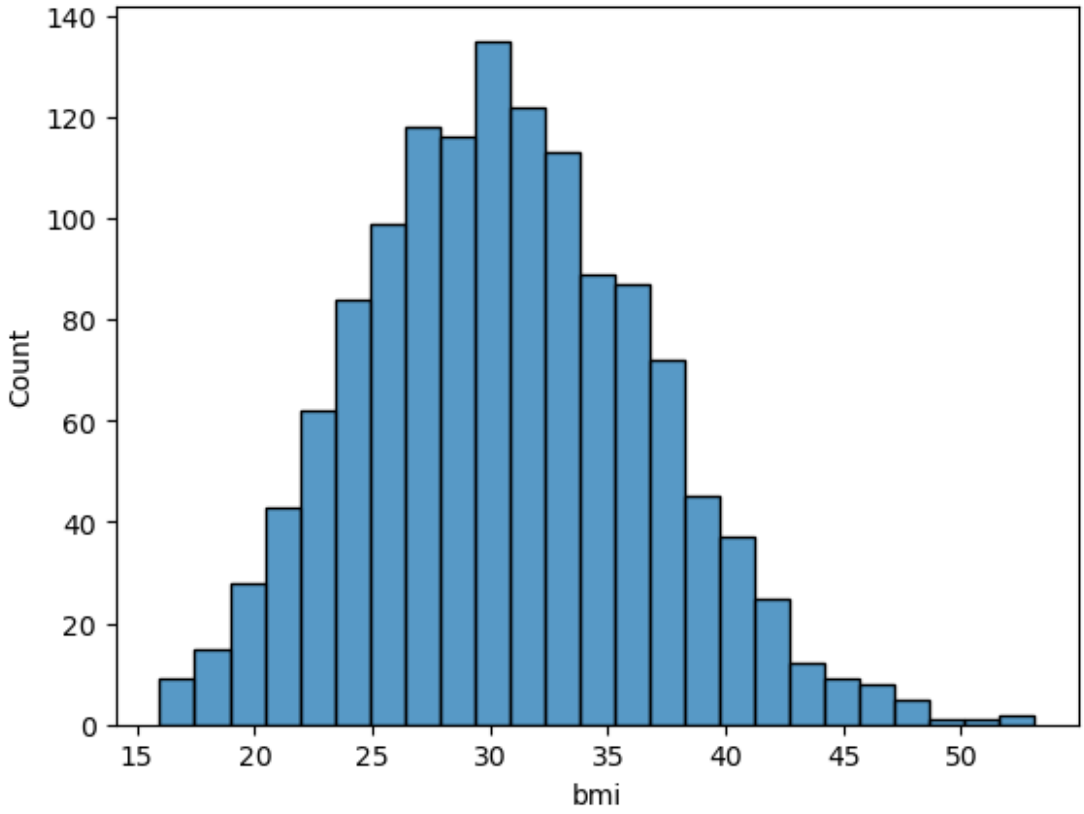
```
      age  is_male     bmi  children  is_smoker      charges  \
0      19        0  27.900         0          1  16884.92400
1      18        1  33.770         1          0   1725.55230
2      28        1  33.000         3          0   4449.46200
3      33        1  22.705         0          0  21984.47061
4      32        1  28.880         0          0   3866.85520
...   ...      ...     ...       ...        ...          ...
1333   50        1  30.970         3          0  10600.54830
1334   18        0  31.920         0          0   2205.98080
1335   18        0  36.850         0          0   1629.83350
```

```
1336   21         0  25.800         0          0   2007.94500
1337   61         0  29.070         0          1  29141.36030

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

[1337 rows x 10 columns]
```

FEATURE Engineering and extraction

```
sns.histplot(df_clean['bmi'])
```

```
<Axes: xlabel='bmi', ylabel='Count'>
```




```
df_clean['bmi_category']=pd.cut(

    df_clean['bmi'],bins=[0,18.5,24.9,29.9,float('inf')],
    labels=['Underweights','Normal','Overweight','obese']

)

df_clean
```

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

      region_southwest  region_southeast  region_northwest
region_northeast  \
0                    1                 0                 0
```

```
0
1                    0                   1                   0
0
2                    0                   1                   0
0
3                    0                   0                   1
0
4                    0                   0                   1
0
...                ...                 ...                 ...
...
1333                 0                   0                   1
0
1334                 0                   0                   0
1
1335                 0                   1                   0
0
1336                 1                   0                   0
0
1337                 0                   0                   1
0

     bmi_category
0      Overweight
1           obese
2           obese
3          Normal
4      Overweight
...              ...
1333        obese
1334        obese
1335        obese
1336   Overweight
1337   Overweight

[1337 rows x 11 columns]

df_clean['age_category']=pd.cut(

    df_clean['age'],bins=[13,18,39,59,float('inf')],
    labels=['Teen','Adult','Middle-aged','Seniors']

)

df_clean

      age  is_male     bmi  children  is_smoker      charges  \
0      19        0  27.900         0          1  16884.92400
1      18        1  33.770         1          0   1725.55230
2      28        1  33.000         3          0   4449.46200
```

```
3      33         1  22.705         0          0  21984.47061
4      32         1  28.880         0          0   3866.85520
...   ...       ...     ...       ...        ...          ...
1333   50         1  30.970         3          0  10600.54830
1334   18         0  31.920         0          0   2205.98080
1335   18         0  36.850         0          0   1629.83350
1336   21         0  25.800         0          0   2007.94500
1337   61         0  29.070         0          1  29141.36030

      region_southwest  region_southeast  region_northwest
region_northeast  \
0                    1                 0                 0
0
1                    0                 1                 0
0
2                    0                 1                 0
0
3                    0                 0                 1
0
4                    0                 0                 1
0
...                ...               ...               ...
...
1333                 0                 0                 1
0
1334                 0                 0                 0
1
1335                 0                 1                 0
0
1336                 1                 0                 0
0
1337                 0                 0                 1
0

     bmi_category age_category
0      Overweight        Adult
1           obese         Teen
2           obese        Adult
3          Normal        Adult
4      Overweight        Adult
...           ...          ...
1333        obese  Middle-aged
1334        obese         Teen
1335        obese         Teen
1336   Overweight        Adult
1337   Overweight      Seniors

[1337 rows x 12 columns]
```

```
df_clean['age_category'].value_counts()
```

```
age_category
Adult          604
Middle-aged    550
Seniors        114
Teen            69
Name: count, dtype: int64
```

```
df_clean=pd.get_dummies(df_clean,columns=['bmi_category'],drop_first=True)
```

```
df_clean=pd.get_dummies(df_clean,columns=['age_category'],drop_first=True)
```

```
df_clean
```

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

      region_southwest  region_southeast  region_northwest
region_northeast  \
0                    1                 0                 0
0
1                    0                 1                 0
0
2                    0                 1                 0
0
3                    0                 0                 1
0
4                    0                 0                 1
0
...                ...               ...               ...
...
1333                 0                 0                 1
0
1334                 0                 0                 0
1
1335                 0                 1                 0
0
1336                 1                 0                 0
0
```

```
1337                  0                  0                  1
0

      bmi_category_Normal  bmi_category_Overweight  bmi_category_obese
\
0                   False                    True               False

1                   False                   False                True

2                   False                   False                True

3                    True                   False               False

4                   False                    True               False

...                   ...                     ...                 ...

1333                False                   False                True

1334                False                   False                True

1335                False                   False                True

1336                False                    True               False

1337                False                    True               False


      age_category_Adult  age_category_Middle-aged
age_category_Seniors
0                   True                   False
False
1                  False                   False
False
2                   True                   False
False
3                   True                   False
False
4                   True                   False
False
...                  ...                     ...                  .
..
1333               False                    True
False
1334               False                   False
False
1335               False                   False
False
1336                True                   False
False
```

```
1337               False                       False
True

[1337 rows x 16 columns]
```

```
df_clean=df_clean.astype(int)
```

```
df_clean
```

```
      age  is_male  bmi  children  is_smoker  charges
region_southwest  \
0      19        0   27         0          1    16884
1
1      18        1   33         1          0     1725
0
2      28        1   33         3          0     4449
0
3      33        1   22         0          0    21984
0
4      32        1   28         0          0     3866
0
...   ...      ...  ...       ...        ...      ...                ..
.
1333   50        1   30         3          0    10600
0
1334   18        0   31         0          0     2205
0
1335   18        0   36         0          0     1629
0
1336   21        0   25         0          0     2007
1
1337   61        0   29         0          1    29141
0

      region_southeast  region_northwest  region_northeast  \
0                    0                 0                 0
1                    1                 0                 0
2                    1                 0                 0
3                    0                 1                 0
4                    0                 1                 0
...                ...               ...               ...
1333                 0                 1                 0
1334                 0                 0                 1
1335                 1                 0                 0
1336                 0                 0                 0
1337                 0                 1                 0

      bmi_category_Normal  bmi_category_Overweight  bmi_category_obese
\
0                       0                        1                   0
```

```
1                          0                          0                    1
2                          0                          0                    1
3                          1                          0                    0
4                          0                          1                    0
...                      ...                        ...                  ...
1333                       0                          0                    1
1334                       0                          0                    1
1335                       0                          0                    1
1336                       0                          1                    0
1337                       0                          1                    0

      age_category_Adult  age_category_Middle-aged
age_category_Seniors
0                       1                          0
0
1                       0                          0
0
2                       1                          0
0
3                       1                          0
0
4                       1                          0
0
...                   ...                        ...                    .
..
1333                    0                          1
0
1334                    0                          0
0
1335                    0                          0
0
1336                    1                          0
0
1337                    0                          0
1

[1337 rows x 16 columns]
```

Feature Scaling

```
from sklearn.preprocessing import StandardScaler

cols=['age','bmi','children']

scaler=StandardScaler()
df_clean[cols]=scaler.fit_transform(df_clean[cols])

df_clean
```

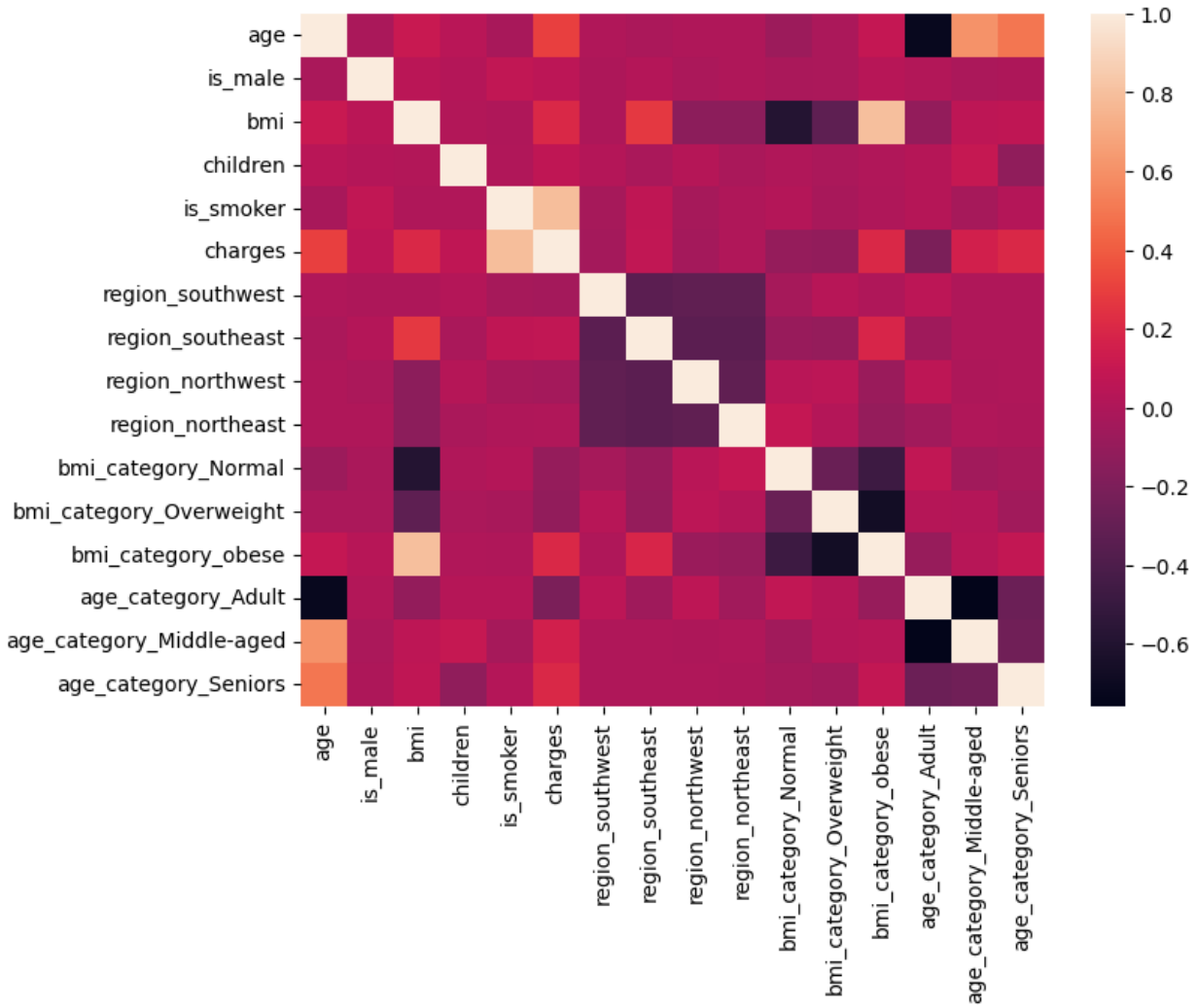
```
           age  is_male       bmi  children  is_smoker  charges  \
0    -1.440418        0 -0.517949 -0.909234          1    16884
1    -1.511647        1  0.462463 -0.079442          0     1725
2    -0.799350        1  0.462463  1.580143          0     4449
3    -0.443201        1 -1.334960 -0.909234          0    21984
4    -0.514431        1 -0.354547 -0.909234          0     3866
...        ...      ...       ...       ...        ...      ...
1333  0.767704        1 -0.027743  1.580143          0    10600
1334 -1.511647        0  0.135659 -0.909234          0     2205
1335 -1.511647        0  0.952670 -0.909234          0     1629
1336 -1.297958        0 -0.844753 -0.909234          0     2007
1337  1.551231        0 -0.191145 -0.909234          1    29141

      region_southwest  region_southeast  region_northwest
region_northeast  \
0                    1                 0                 0
0
1                    0                 1                 0
0
2                    0                 1                 0
0
3                    0                 0                 1
0
4                    0                 0                 1
0
...                ...               ...               ...
...
1333                 0                 0                 1
0
1334                 0                 0                 0
1
1335                 0                 1                 0
0
1336                 1                 0                 0
0
1337                 0                 0                 1
0

      bmi_category_Normal  bmi_category_Overweight  bmi_category_obese
\
```

```
0                         0                         1                         0
1                         0                         0                         1
2                         0                         0                         1
3                         1                         0                         0
4                         0                         1                         0
...                     ...                       ...                       ...
1333                      0                         0                         1
1334                      0                         0                         1
1335                      0                         0                         1
1336                      0                         1                         0
1337                      0                         1                         0


      age_category_Adult  age_category_Middle-aged
age_category_Seniors
0                         1                         0
0
1                         0                         0
0
2                         1                         0
0
3                         1                         0
0
4                         1                         0
0
...                     ...                       ...                       .
..
1333                      0                         1
0
1334                      0                         0
0
1335                      0                         0
0
1336                      1                         0
0
1337                      0                         0
1

[1337 rows x 16 columns]
```

Feature Selection

```
plt.figure(figsize=(8,6))
sns.heatmap(df_clean.corr(numeric_only=True))
```

```
<Axes: >
```




```
corr_with_charges = df_clean.corr()['charges']
corr_with_charges=corr_with_charges.sort_values(ascending=False)

corr_with_charges[corr_with_charges>0]
```

```
charges                    1.000000
is_smoker                  0.787234
age                        0.298309
age_category_Seniors       0.200975
bmi_category_obese         0.197660
bmi                        0.196236
```

```
age_category_Middle-aged    0.148667
region_southeast            0.073577
children                    0.067390
is_male                     0.058046
region_northeast            0.005946
Name: charges, dtype: float64
```

```
df_clean.columns
```

```
Index(['age', 'is_male', 'bmi', 'children', 'is_smoker', 'charges',
       'region_southwest', 'region_southeast', 'region_northwest',
       'region_northeast', 'bmi_category_Normal',
'bmi_category_Overweight',
       'bmi_category_obese', 'age_category_Adult',
'age_category_Middle-aged',
       'age_category_Seniors'],
      dtype='object')
```

```
from scipy.stats import pearsonr

# ---------------------------------
# Pearson Correlation Calculation
# ---------------------------------

# List of features to check against target
selected_features = ['age', 'is_male', 'bmi', 'children', 'is_smoker',
       'region_southwest', 'region_southeast', 'region_northwest',
       'region_northeast', 'bmi_category_Normal',
'bmi_category_Overweight',
       'bmi_category_obese', 'age_category_Adult',
'age_category_Middle-aged',
       'age_category_Seniors']

correlations = {
    feature: pearsonr(df_clean[feature], df_clean['charges'])[0]
    for feature in selected_features
}
correlation_df = pd.DataFrame(list(correlations.items()),
columns=['Feature', 'Pearson Correlation'])
correlation_df.sort_values(by='Pearson Correlation', ascending=False)
```

	Feature	Pearson Correlation
4	is_smoker	0.787234
0	age	0.298309
14	age_category_Seniors	0.200975
11	bmi_category_obese	0.197660
2	bmi	0.196236
13	age_category_Middle-aged	0.148667
6	region_southeast	0.073577
3	children	0.067390

```
1                     is_male             0.058046
8           region_northeast              0.005946
7           region_northwest             -0.038695
5           region_southwest             -0.043637
9        bmi_category_Normal             -0.105656
10   bmi_category_Overweight             -0.118280
12        age_category_Adult             -0.206731
```

```
cat_features=[ 'is_male', 'is_smoker',
       'region_southwest', 'region_southeast', 'region_northwest',
       'region_northeast', 'bmi_category_Normal',
'bmi_category_Overweight',
       'bmi_category_obese', 'age_category_Adult',
'age_category_Middle-aged',
       'age_category_Seniors']

from scipy.stats import chi2_contingency

alpha = 0.05

df_clean['charges_bin'] = pd.qcut(df_clean['charges'], q=4,
labels=False)
chi2_results = {}

for col in cat_features:
    contingency = pd.crosstab(df_clean[col], df_clean['charges_bin'])
    chi2_stat, p_val, _, _ = chi2_contingency(contingency)
    decision = 'Reject Null (Keep Feature)' if p_val < alpha else
'Accept Null (Drop Feature)'
    chi2_results[col] = {
        'chi2_statistic': chi2_stat,
        'p_value': p_val,
        'Decision': decision
    }

chi2_df = pd.DataFrame(chi2_results).T
chi2_df = chi2_df.sort_values(by='p_value')
chi2_df
```

```
                          chi2_statistic   p_value
Decision
is_smoker                     848.219178       0.0  Reject Null (Keep
Feature)
age_category_Adult            407.358116       0.0  Reject Null (Keep
Feature)
age_category_Middle-aged      352.360041       0.0  Reject Null (Keep
Feature)
age_category_Seniors          161.971585       0.0  Reject Null (Keep
Feature)
region_southeast               15.998167  0.001135  Reject Null (Keep
```

```
Feature)
is_male                        10.258784   0.01649  Reject Null (Keep
Feature)
bmi_category_obese              7.654464   0.05372  Accept Null (Drop
Feature)
region_northeast                6.438442  0.092122  Accept Null (Drop
Feature)
region_southwest                5.091893  0.165191  Accept Null (Drop
Feature)
bmi_category_Normal             4.263673  0.234364  Accept Null (Drop
Feature)
bmi_category_Overweight         4.201575  0.240504  Accept Null (Drop
Feature)
region_northwest                 1.13424  0.768815  Accept Null (Drop
Feature)
```

```
df_clean=df_clean[['age','is_male','bmi','is_smoker','age_category_Adu
lt','age_category_Middle-
aged','age_category_Seniors','region_southeast']]
```

```
df_clean
```

```
           age  is_male       bmi  is_smoker  age_category_Adult  \
0    -1.440418        0 -0.517949          1                   1
1    -1.511647        1  0.462463          0                   0
2    -0.799350        1  0.462463          0                   1
3    -0.443201        1 -1.334960          0                   1
4    -0.514431        1 -0.354547          0                   1
...        ...      ...       ...        ...                 ...
1333  0.767704        1 -0.027743          0                   0
1334 -1.511647        0  0.135659          0                   0
1335 -1.511647        0  0.952670          0                   0
1336 -1.297958        0 -0.844753          0                   1
1337  1.551231        0 -0.191145          1                   0

      age_category_Middle-aged  age_category_Seniors  region_southeast

0                            0                     0                 0

1                            0                     0                 1

2                            0                     0                 1

3                            0                     0                 0

4                            0                     0                 0

...                        ...                   ...               ...

1333                         1                     0                 0
```

```
1334                              0                              0                              0
1335                              0                              0                              1
1336                              0                              0                              0
1337                              0                              1                              0

[1337 rows x 8 columns]
#df_clean.to_csv("INSURANCE_CLEAN(Ready).csv",index=False)
```