

```
In [15]: import pandas as pd
```

```
In [16]: df = pd.read_csv('data.csv')
df
```

```
Out[16]:
```

	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
...	...	...	...	...	...	...	...
1333	50	male	30.970	3	no	northwest	10600.54830
1334	18	female	31.920	0	no	northeast	2205.98080
1335	18	female	36.850	0	no	southeast	1629.83350
1336	21	female	25.800	0	no	southwest	2007.94500
1337	61	female	29.070	0	yes	northwest	29141.36030

1338 rows × 7 columns

```
In [17]: df['sex'] = df['sex'].apply(lambda x: 1 if x == 'male' else 0)
df['smoker'] = df['smoker'].apply(lambda x: 1 if x == 'yes' else 0)
```

```
In [23]: df = df.join(pd.get_dummies(df.region, dtype=int)).drop('region' , axis=1)
```

```
In [24]: df
```

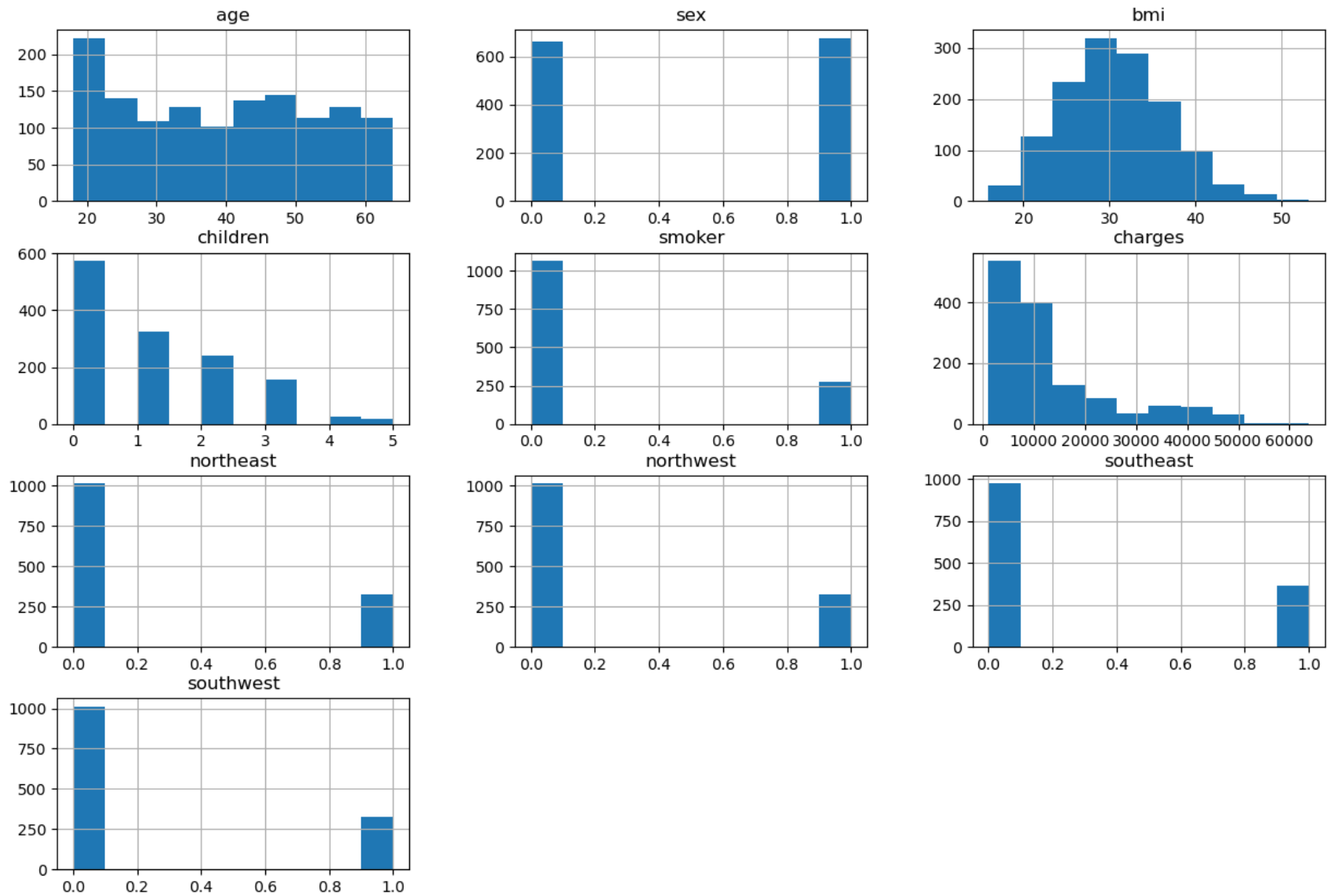
Out [24]:

	age	sex	bmi	children	smoker	charges	northeast	northwest	southeast	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	1	0	0
...	...	...	...	...	...	...	...	...	...	...
1333	50	1	30.970	3	0	10600.54830	0	1	0	0
1334	18	0	31.920	0	0	2205.98080	1	0	0	0
1335	18	0	36.850	0	0	1629.83350	0	0	1	0
1336	21	0	25.800	0	0	2007.94500	0	0	0	1
1337	61	0	29.070	0	1	29141.36030	0	1	0	0

1338 rows x 10 columns

In [26]: `df.hist(figsize=(15,10))`

Out [26]: array([[<Axes: title={'center': 'age'}>, <Axes: title={'center': 'sex'}>,
 <Axes: title={'center': 'bmi'}>],
 [<Axes: title={'center': 'children'}>,
 <Axes: title={'center': 'smoker'}>,
 <Axes: title={'center': 'charges'}>],
 [<Axes: title={'center': 'northeast'}>,
 <Axes: title={'center': 'northwest'}>,
 <Axes: title={'center': 'southeast'}>],
 [<Axes: title={'center': 'southwest'}>, <Axes: >, <Axes: >]],
 dtype=object)



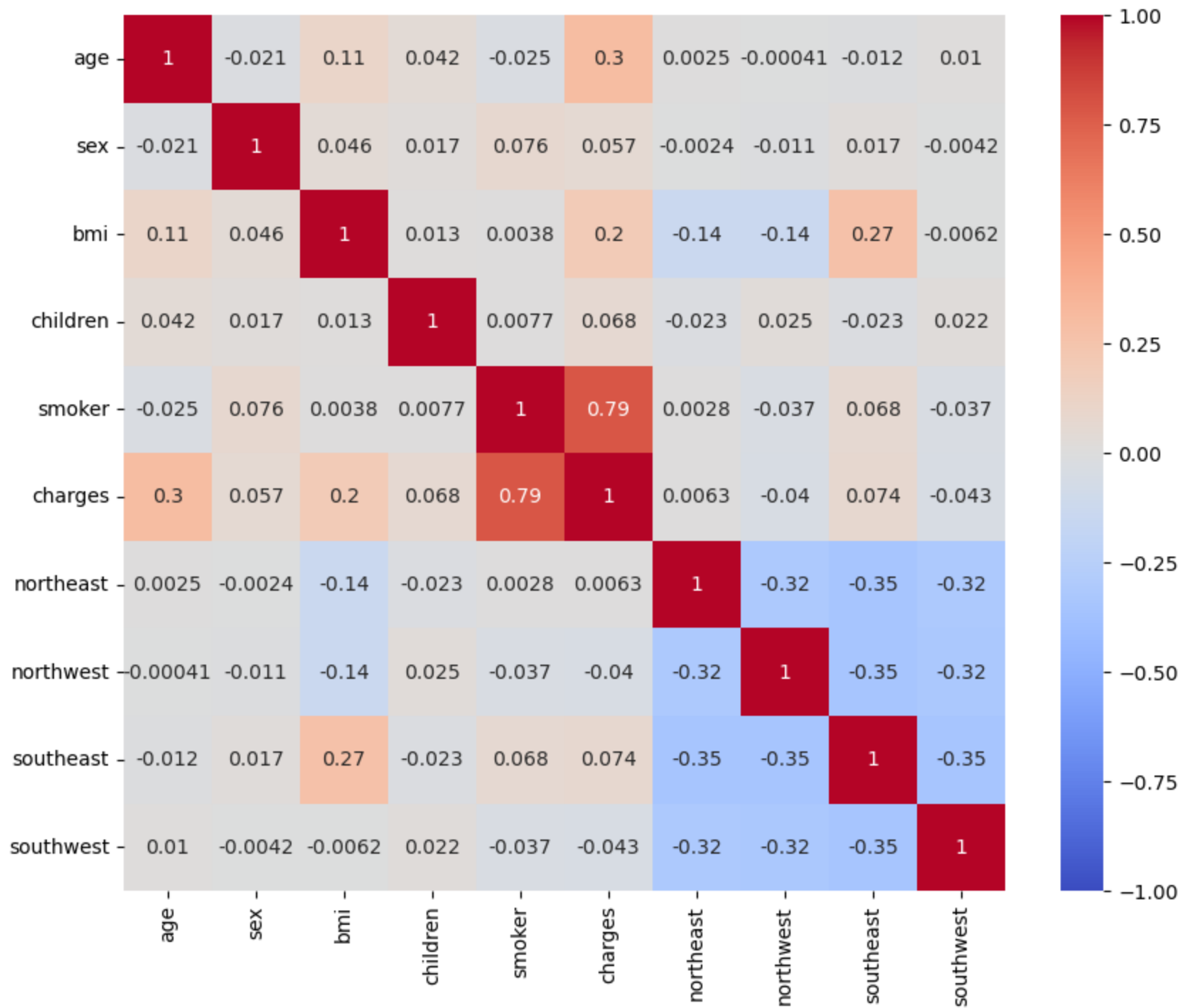
```
In [27]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 10 columns):
#   Column      Non-Null Count  Dtype
---  -
0   age         1338 non-null   int64
1   sex         1338 non-null   int64
2   bmi         1338 non-null   float64
3   children    1338 non-null   int64
4   smoker      1338 non-null   int64
5   charges     1338 non-null   float64
6   northeast   1338 non-null   int64
7   northwest   1338 non-null   int64
8   southeast   1338 non-null   int64
9   southwest   1338 non-null   int64
dtypes: float64(2), int64(8)
memory usage: 104.7 KB
```

```
In [32]: import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(10,8))
sns.heatmap(df.corr(), annot=True, cmap='coolwarm', vmin=-1, vmax=1)
```

```
Out[32]: <Axes: >
```



```
In [73]: import statsmodels.api as sm
```

```
In [74]: X = df[['age', 'bmi', 'smoker', 'southeast']]  
y = df['charges']
```

```
In [75]: X = sm.add_constant(X)
```

```
In [76]: model = sm.OLS(y, X).fit()
```

```
In [78]: model.summary()
```

Out [78]:

OLS Regression Results

Dep. Variable:	charges	R-squared:	0.748
Model:	OLS	Adj. R-squared:	0.747
Method:	Least Squares	F-statistic:	988.9
Date:	Tue, 10 Jun 2025	Prob (F-statistic):	0.00
Time:	21:31:15	Log-Likelihood:	-13556.
No. Observations:	1338	AIC:	2.712e+04
Df Residuals:	1333	BIC:	2.715e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
const	-1.187e+04	944.659	-12.560	0.000	-1.37e+04	-1e+04
age	258.7708	11.938	21.677	0.000	235.352	282.190
bmi	334.8987	28.558	11.727	0.000	278.875	390.923
smoker	2.387e+04	413.626	57.706	0.000	2.31e+04	2.47e+04
southeast	-613.7877	389.780	-1.575	0.116	-1378.438	150.862

Omnibus:	301.049	Durbin-Watson:	2.079
Prob(Omnibus):	0.000	Jarque-Bera (JB):	714.025
Skew:	1.218	Prob(JB):	8.94e-156
Kurtosis:	5.622	Cond. No.	292.

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

```
In [ ]: age = int(input("What is your age?"))
bmi = float(input("What is your BMI?"))
smoker_input = input("Are you a smoker? (yes/no)").strip().lower()
region_input = input("What region do you reside? (e.g., southeast, northwest, northeast, southwest)").strip()

smoker = 1 if smoker_input == 'yes' else 0
southeast = 1 if region_input == 'southeast' else 0

new_user_df = pd.DataFrame({
    'const': [1],
    'age': [age],
    'bmi': [bmi],
    'smoker': [smoker],
    'southeast': [southeast]
})

# Predict charges
predicted_charges = model.predict(new_user_df)[0]
```

```
In [ ]:
```