Medical Expenses Prediction

About the Data:

The medical insurance dataset encompasses various factors influencing medical expenses, such as age, sex, BMI, smoking status, number of children, and region. This dataset serves as a foundation for training machine learning models capable of forecasting medical expenses for new policyholders.

Goal of Project:

The goal of this project is to build two models using this data:

- A descriptive model for understanding what features significantly effect medical costs, and to what extend they do so.
- A predictive model that can take input and predict medical costs with high accuracy.

The predictive model will then be deployed locally using joblib, fastapi, and docker.

libraries

In [1]:
```
import pandas as pd
import numpy as np

import matplotlib.pyplot as plt
import seaborn as sns

import statsmodels.api as sm
from scipy.stats import shapiro

from sklearn.ensemble import RandomForestRegressor
from sklearn.pipeline import Pipeline
from sklearn.model_selection import GridSearchCV, train_test_split, cross_val_score
from sklearn.metrics import r2_score, mean_squared_error
```

data

In [2]:
```
df = pd.read_csv('medical_insurance.csv')
```

In [3]:
```
df.head()
```

Out[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

Data Prep

shape, missing, duplicates

In [4]:
```
print(f'shape:{df.shape}\n')
print(f'missing:\n{df.isna().sum()}\n')
print(f'duplicates:{df.duplicated().sum()}')
```

```
shape:(2772, 7)

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

duplicates:1435
```

In [5]:

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

> duplicates dropped

---

unique values in columns

In [6]:

```
for col in df.columns:
    print(col, df[col].unique())
```

```
age [19 18 28 33 32 31 46 37 60 25 62 23 56 27 52 30 34 59 63 55 22 26 35 24
 41 38 36 21 48 40 58 53 43 64 20 61 44 57 29 45 54 49 47 51 42 50 39]
sex ['female' 'male']
bmi [27.9   33.77  33.    22.705 28.88  25.74  33.44  27.74  29.83  25.84
 26.22  26.29  34.4   39.82  42.13  24.6   30.78  23.845 40.3   35.3
 36.005 32.4   34.1   31.92  28.025 27.72  23.085 32.775 17.385 36.3
 35.6   26.315 28.6   28.31  36.4   20.425 32.965 20.8   36.67  39.9
 26.6   36.63  21.78  30.8   37.05  37.3   38.665 34.77  24.53  35.2
 35.625 33.63  28.    34.43  28.69  36.955 31.825 31.68  22.88  37.335
 27.36  33.66  24.7   25.935 22.42  28.9   39.1   36.19  23.98  24.75
 28.5   28.1   32.01  27.4   34.01  29.59  35.53  39.805 26.885 38.285
 37.62  41.23  34.8   22.895 31.16  27.2   26.98  39.49  24.795 31.3
 38.28  19.95  19.3   31.6   25.46  30.115 29.92  27.5   28.4   30.875
 27.94  35.09  29.7   35.72  32.205 28.595 49.06  27.17  23.37  37.1
 23.75  28.975 31.35  33.915 28.785 28.3   37.4   17.765 34.7   26.505
 22.04  35.9   25.555 28.05  25.175 31.9   36.    32.49  25.3   29.735
 38.83  30.495 37.73  37.43  24.13  37.145 39.52  24.42  27.83  36.85
 39.6   29.8   29.64  28.215 37.    33.155 18.905 41.47  30.3   15.96
 33.345 37.7   27.835 29.2   26.41  30.69  41.895 30.9   32.2   32.11
 31.57  26.2   30.59  32.8   18.05  39.33  32.23  24.035 36.08  22.3
 26.4   31.8   26.73  23.1   23.21  33.7   33.25  24.64  33.88  38.06
 41.91  31.635 36.195 17.8   24.51  22.22  38.39  29.07  22.135 26.8
 30.02  35.86  20.9   17.29  34.21  25.365 40.15  24.415 25.2   26.84
 24.32  42.35  19.8   32.395 30.2   29.37  34.2   27.455 27.55  20.615
 24.3   31.79  21.56  28.12  40.565 27.645 31.2   26.62  48.07  36.765
 33.4   45.54  28.82  22.99  27.7   25.41  34.39  22.61  37.51  38.
 33.33  34.865 33.06  35.97  31.4   25.27  40.945 34.105 36.48  33.8
 36.7   36.385 34.5   32.3   27.6   29.26  35.75  23.18  25.6   35.245
 43.89  20.79  30.5   21.7   21.89  24.985 32.015 30.4   21.09  22.23
 32.9   24.89  31.46  17.955 30.685 43.34  39.05  30.21  31.445 19.855
 31.02  38.17  20.6   47.52  20.4   38.38  24.31  23.6   21.12  30.03
 17.48  20.235 17.195 23.9   35.15  35.64  22.6   39.16  27.265 29.165
 16.815 33.1   26.9   33.11  31.73  46.75  29.45  32.68  33.5   43.01
 36.52  26.695 25.65  29.6   38.6   23.4   46.53  30.14  30.    38.095
 28.38  28.7   33.82  24.09  32.67  25.1   32.56  41.325 39.5   34.3
 31.065 21.47  25.08  43.4   25.7   27.93  39.2   26.03  30.25  28.93
 35.7   35.31  31.    44.22  26.07  25.8   39.425 40.48  38.9   47.41
 35.435 46.7   46.2   21.4   23.8   44.77  32.12  29.1   37.29  43.12
 36.86  34.295 23.465 45.43  23.65  20.7   28.27  35.91  29.    19.57
 31.13  21.85  40.26  33.725 29.48  32.6   37.525 23.655 37.8   19.
 21.3   33.535 42.46  38.95  36.1   29.3   39.7   38.19  42.4   34.96
 42.68  31.54  29.81  21.375 40.81  17.4   20.3   18.5   26.125 41.69
 24.1   36.2   40.185 39.27  34.87  44.745 29.545 23.54  40.47  40.66
 36.6   35.4   27.075 28.405 21.755 40.28  30.1   32.1   23.7   35.5
 29.15  27.    37.905 22.77  22.8   34.58  27.1   19.475 26.7   34.32
 24.4   41.14  22.515 41.8   26.18  42.24  26.51  35.815 41.42  36.575
 42.94  21.01  24.225 17.67  31.5   31.1   32.78  32.45  50.38  47.6
 25.4   29.9   43.7   24.86  28.8   29.5   29.04  38.94  44.    20.045
```

```
 40.92  35.1   29.355 32.585 32.34  39.8   24.605 33.99  28.2   25.
 33.2   23.2   20.1   32.5   37.18  46.09  39.93  35.8   31.255 18.335
 42.9   26.79  39.615 25.9   25.745 28.16  23.56  40.5   35.42  39.995
 34.675 20.52  23.275 36.29  32.7   19.19  20.13  23.32  45.32  34.6
 18.715 21.565 23.    37.07  52.58  42.655 21.66  32.    18.3   47.74
 22.1   19.095 31.24  29.925 20.35  25.85  42.75  18.6   23.87  45.9
 21.5   30.305 44.88  41.1   40.37  28.49  33.55  40.375 27.28  17.86
 33.3   39.14  21.945 24.97  23.94  34.485 21.8   23.3   36.96  21.28
 29.4   27.3   37.9   37.715 23.76  25.52  27.61  27.06  39.4   34.9
 22.    30.36  27.8   53.13  39.71  32.87  44.7   30.97 ]
children [0 1 3 2 5 4]
smoker ['yes' 'no']
region ['southwest' 'southeast' 'northwest' 'northeast']
charges [16884.924   1725.5523  4449.462  ...  1629.8335  2007.945  29141.3603]
```

> There doesn't appear to be any errors

---

data types

In [7]:
```python
df.dtypes
```

Out[7]:
```
age           int64
sex          object
bmi         float64
children      int64
smoker       object
region       object
charges     float64
dtype: object
```

dummy variables

In [8]:
```python
df_dummies = pd.get_dummies(df, columns=['sex','smoker','region'], drop_first=True)
```

In [9]:
```python
df_dummies.head()
```

Out[9]:

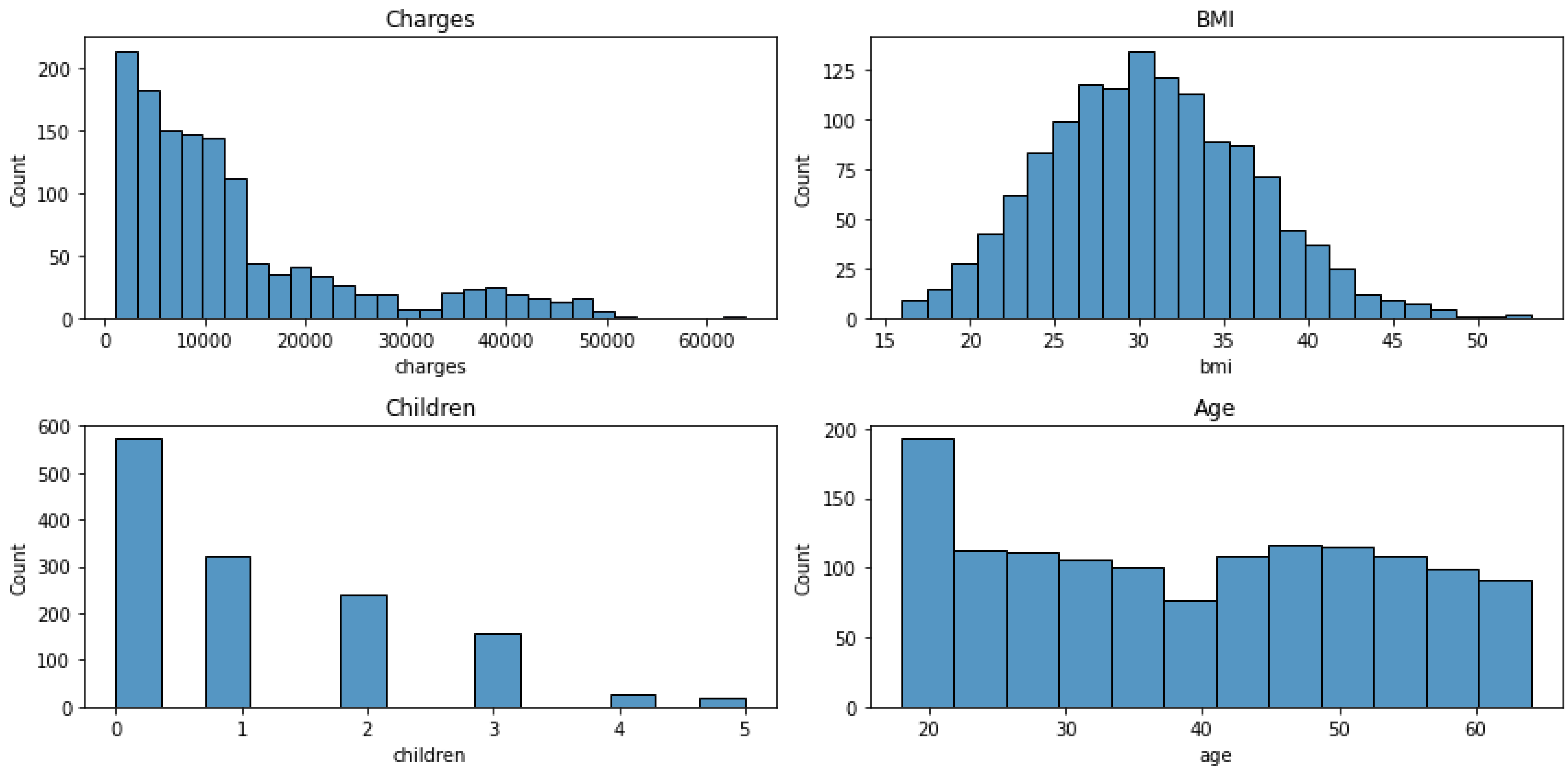
	age	bmi	children	charges	sex_male	smoker_yes	region_northwest	region_southeast	region_southwest
0	19	27.900	0	16884.92400	False	True	False	False	True
1	18	33.770	1	1725.55230	True	False	False	True	False
2	28	33.000	3	4449.46200	True	False	False	True	False
3	33	22.705	0	21984.47061	True	False	True	False	False
4	32	28.880	0	3866.85520	True	False	True	False	False

EDA

continuous distributions

In [10]:

```
#
cols   = ['charges','bmi','children','age']

fig, ((ax1, ax2), (ax3, ax4)) = plt.subplots(2,2, figsize=(12,6))
sns.histplot(df_dummies.charges, ax=ax1).set_title('Charges')
sns.histplot(df_dummies.bmi, ax=ax2).set_title('BMI')
sns.histplot(df_dummies.children, ax=ax3).set_title('Children')
sns.histplot(df_dummies.age, ax=ax4).set_title('Age')

plt.tight_layout()
```

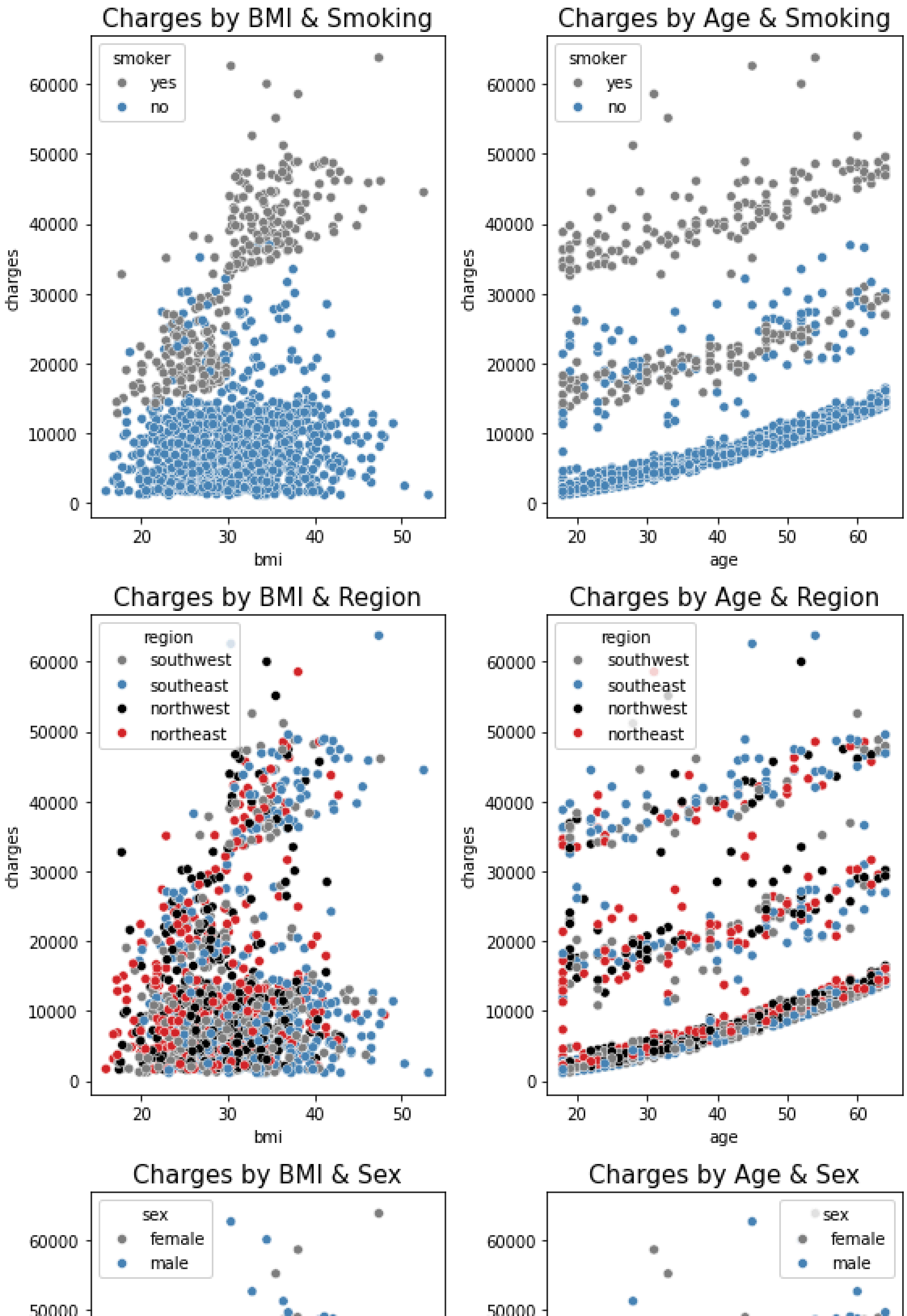


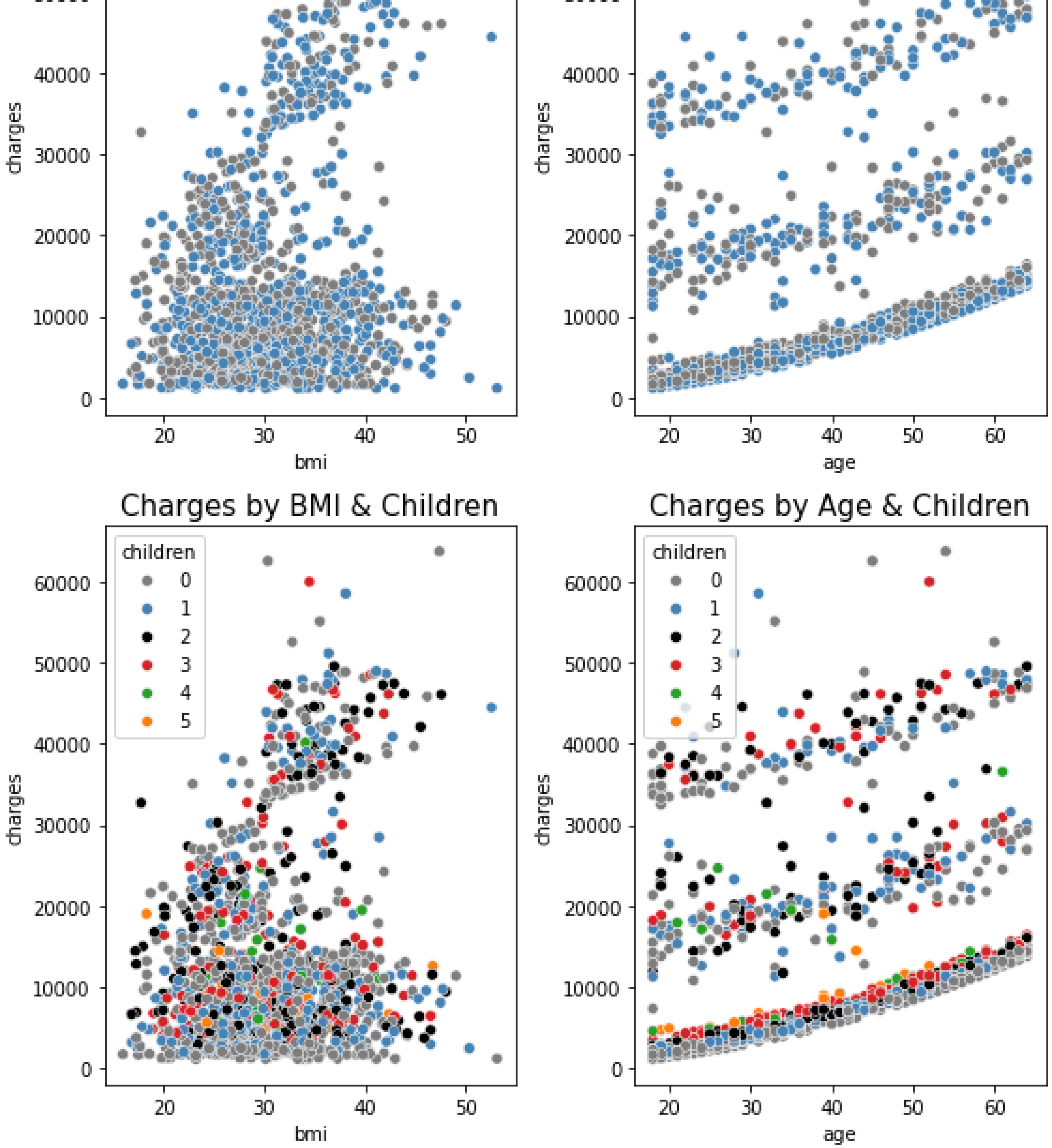
> The target variable 'charges' is highly skewed and may need to be capped or log transformed

possible interaction terms

In [11]:
```
#
palette = 'grey','steelblue','black','tab:red', 'tab:green', 'tab:orange'
palette2 = ['grey','steelblue','black','tab:red', 'tab:green', 'tab:orange']

fig, ((ax1,ax2),(ax3,ax4),(ax5,ax6),(ax7,ax8)) = plt.subplots(4,2, figsize=(8,20))

sns.scatterplot(data=df, x='bmi', y='charges', ax=ax1, hue='smoker', palette=palette).set_title('Charges by BMI & Smoking', size=1
sns.scatterplot(data=df, x='age', y='charges', ax=ax2, hue='smoker', palette=palette).set_title('Charges by Age & Smoking',size=15

sns.scatterplot(data=df, x='bmi', y='charges', ax=ax3, hue='region', palette=palette).set_title('Charges by BMI & Region',size=15)
sns.scatterplot(data=df, x='age', y='charges', ax=ax4, hue='region', palette=palette).set_title('Charges by Age & Region',size=15)
```

```
sns.scatterplot(data=df, x='bmi', y='charges', ax=ax5, hue='sex', palette=palette).set_title('Charges by BMI & Sex',size=15)
sns.scatterplot(data=df, x='age', y='charges', ax=ax6, hue='sex', palette=palette).set_title('Charges by Age & Sex',size=15)

sns.scatterplot(data=df, x='bmi', y='charges', ax=ax7, hue='children', palette=palette2).set_title('Charges by BMI & Children',siz
sns.scatterplot(data=df, x='age', y='charges', ax=ax8, hue='children', palette=palette2).set_title('Charges by Age & Children',siz

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


Charges by BMI & Smoking
smoker
yes
no
charges
bmi
Charges by Age & Smoking
smoker
yes
no
charges
age
Charges by BMI & Region
region
southwest
southeast
northwest
northeast
charges
bmi
Charges by Age & Region
region
southwest
southeast
northwest
northeast
charges
age
Charges by BMI & Sex
sex
female
male
Charges by Age & Sex
sex
female
male

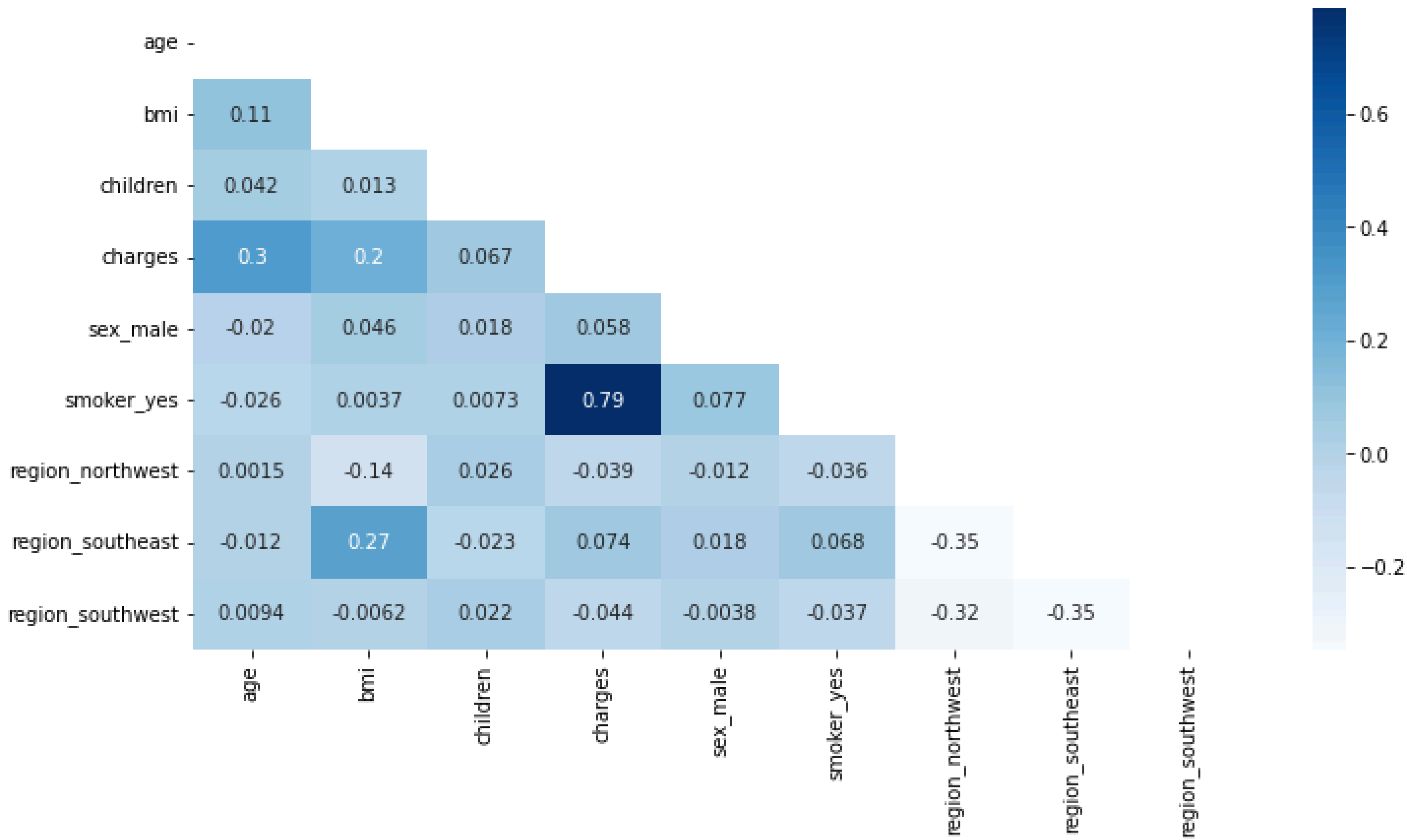


> There apppears to be pretty good separation with bmi & smoking and age & smoking

correlation map

In [12]:
```
#
mask = np.triu(np.ones_like(df_dummies.corr()))

plt.figure(figsize=(12,6))
sns.heatmap(df_dummies.corr(), mask=mask, annot=True, cmap='Blues')
```

Out[12]: <Axes: >



> There is a strong positive correlation between charges and smoking

---

Descriptive Model with Statsmodels

In [13]:
```
X = df_dummies.drop(columns=['charges'])
y = df_dummies.charges
```

In [14]:
```
X_const = sm.add_constant(X).astype(int)
model   = sm.OLS(y, X_const).fit(cov_type='HC3')

print(model.summary())
```

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                charges   R-squared:                       0.751
Model:                            OLS   Adj. R-squared:                  0.749
Method:                 Least Squares   F-statistic:                     297.9
Date:                Sun, 20 Oct 2024   Prob (F-statistic):         5.65e-290
Time:                        11:02:52   Log-Likelihood:                -13538.
No. Observations:                1337   AIC:                         2.709e+04
Df Residuals:                    1328   BIC:                         2.714e+04
Df Model:                           8                                         
Covariance Type:                  HC3                                         
====================================================================================
                       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------
const            -1.174e+04   1035.332    -11.339      0.000   -1.38e+04   -9710.195
age                256.8877     11.987     21.430      0.000     233.393     280.382
bmi                338.0417     31.824     10.622      0.000     275.667     400.416
children           477.5644    131.074      3.643      0.000     220.663     734.465
sex_male          -130.6674    335.233     -0.390      0.697    -787.711     526.376
smoker_yes        2.386e+04    578.212     41.263      0.000     2.27e+04     2.5e+04
region_northwest  -342.7938    487.612     -0.703      0.482   -1298.497     612.909
region_southeast -1042.6351    503.359     -2.071      0.038   -2029.200     -56.070
region_southwest  -969.5839    463.334     -2.093      0.036   -1877.703     -61.465
==============================================================================
Omnibus:                      299.301   Durbin-Watson:                   2.090
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              714.291
Skew:                           1.209   Prob(JB):                    7.83e-156
Kurtosis:                       5.641   Cond. No.                         306.
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)
```

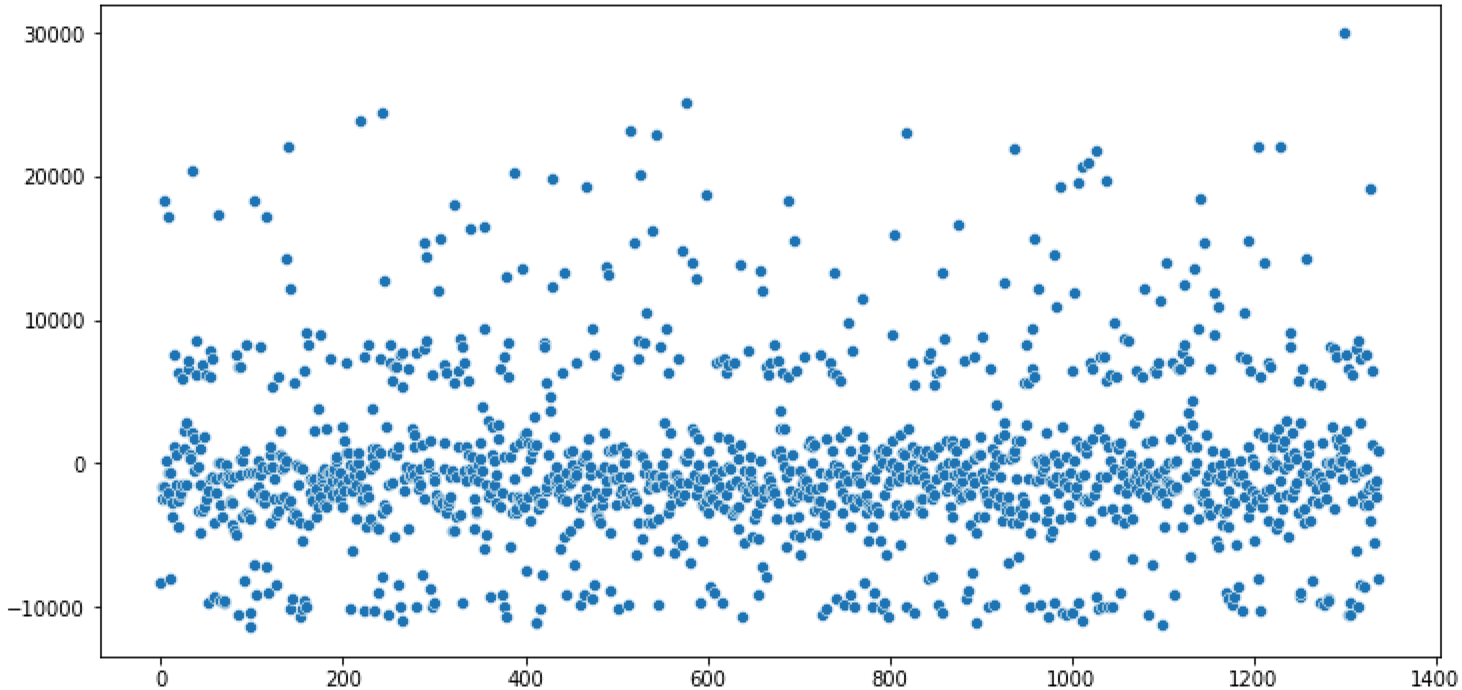
> cov_type='HC3' was used to make the p-values and confidence intervals reliable even though the residuals are non normal

> All of the variables have a significant p-value except for sex_male and region_northwest

---

residuals

In [15]:
```
plt.figure(figsize=(12,6))
sns.scatterplot(model.resid)
plt.show()
print(f'Shapiro-Wilk test p-value:{shapiro(model.resid)[1]}')
```



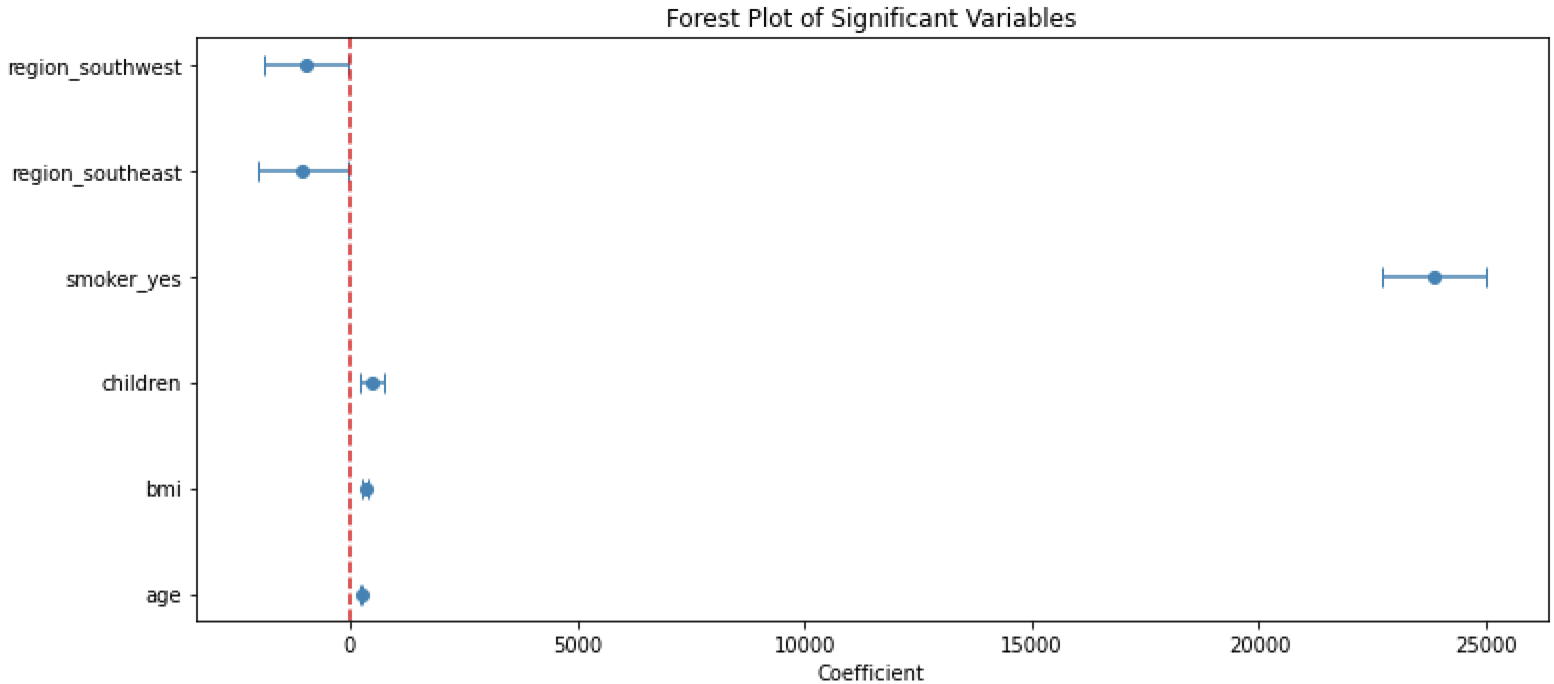

```
Shapiro-Wilk test p-value:9.352632986532134e-29
```

forest plot

In [16]:
```
summary = model.summary2().tables[1]
significant_vars = summary[summary['P>|z|'] < 0.05][1:]
coefficients = significant_vars['Coef.']
conf_int = significant_vars[['[0.025', '0.975]']]
```

In [17]:
```
#
fig, ax = plt.subplots(figsize=(11,5))

ax.errorbar(coefficients, np.arange(len(coefficients)),
            xerr=[coefficients - conf_int['[0.025'], conf_int['0.975]'] - coefficients],
            fmt='o', color='steelblue', capsize=5)
```

```
# Add labels
ax.set_yticks(np.arange(len(coefficients)))
ax.set_yticklabels(significant_vars.index)
ax.axvline(0, color='tab:red', linestyle='--')
ax.set_xlabel('Coefficient')
ax.set_title('Forest Plot of Significant Variables')

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



results

In [18]:
```
significant_vars[['Coef.','[0.025','0.975]']]
```

Out[18]:

	Coef.	[0.025	0.975]
age	256.887746	233.393157	280.382335
bmi	338.041729	275.667015	400.416443
children	477.564421	220.663433	734.465409
smoker_yes	23858.615944	22725.341247	24991.890641
region_southeast	-1042.635127	-2029.200056	-56.070197
region_southwest	-969.583899	-1877.702576	-61.465221

- Each additional year in age results in an increase of about 256 dollars
- Each one unit increase in bmi results in an increase of about 338 dollars
- Each additional child results in an increase of about 477 dollars
- Smokers pay about 23,858 dollars more
- People in the southeast pay about 1,042 dollars less and people in the southwest about 969 dollars less compared to people in the northeast(the upper bound of these confidence intervals are near the zero line indicating they are close to being non significant).

Predictive Model with Sklearn

> Random forest regression is being used to capture nonlinear relationships and interactions

---

train/test split

In [19]:
```
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.2,random_state=1)
```

pipeline and gridsearch

In [20]:
```
pipeline = Pipeline([('forest', RandomForestRegressor(random_state=1))])
```

In [26]:
```
param_grid = {
    'forest__n_estimators': [50, 75, 100],
    'forest__max_depth': [None, 3, 5, 7],
    'forest__min_samples_split': [2, 5, 10],
    'forest__min_samples_leaf': [5, 10, 15],
}

grid_search = GridSearchCV(pipeline, param_grid, cv=5, scoring='r2', n_jobs=-1)
grid_search.fit(X_train, y_train)

print("Best Hyperparameters:", grid_search.best_params_)
```

```
Best Hyperparameters: {'forest__max_depth': 5, 'forest__min_samples_leaf': 10, 'forest__min_samples_split': 2, 'forest__n_estimator
s': 75}
```

In [27]:
```
model = grid_search.best_estimator_
```

metrics

In [28]:
```
y_pred = model.predict(X_test)
y_pred_train = model.predict(X_train)
```

In [29]:
```
print('train r2:', round(r2_score(y_train,y_pred_train),2))
print('test r2:', round(r2_score(y_test,y_pred),2),'\n')

print('train rmse:', round(np.sqrt(mean_squared_error(y_train,y_pred_train)),2))
print('test rmse:', round(np.sqrt(mean_squared_error(y_test,y_pred)),2),'\n')

print('r2 cross val scores:', cross_val_score(model,X,y,cv=5,scoring='r2'))
```

```
train r2: 0.88
test r2: 0.87

train rmse: 4219.64
test rmse: 4163.98

r2 cross val scores: [0.88457966 0.80557239 0.892494   0.84676686 0.87157961]
```

> The r2 and rmse scores look good and the cross validated r2 scores are close enough together to indicate the model is consistent across different subsets of the data.

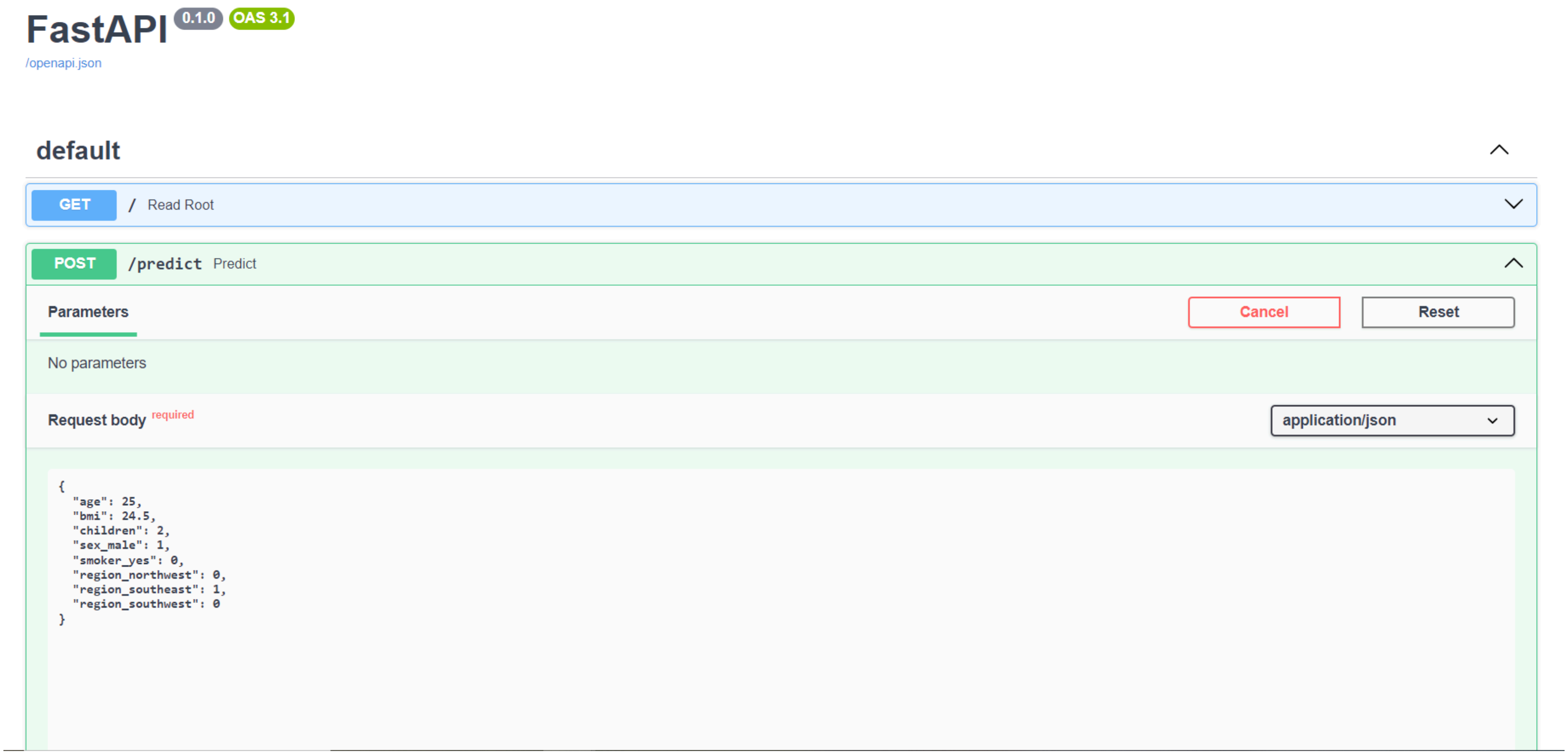
> The range of charges is about 62,648 so the model could be described as on average being off by about 6.6% (given the test rmse).

Save Model

In [30]:
```
import joblib
```

joblib.dump(model, 'medical_expenses.joblib')

Local deployment of a FastAPI app using Docker.


FastAPI 0.1.0 OAS 3.1
/openapi.json
default
GET / Read Root
POST /predict Predict
Parameters
Cancel
Reset
No parameters
Request body required
application/json
{
"age": 25,
"bmi": 24.5,
"children": 2,
"sex_male": 1,
"smoker_yes": 0,
"region_northwest": 0,
"region_southeast": 1,
"region_southwest": 0
}

Curl

```
curl -X 'POST' \
  'http://localhost:8000/predict' \
  -H 'accept: application/json' \
  -H 'Content-Type: application/json' \
  -d '{
  "age": 25,
  "bmi": 24.5,
  "children": 2,
  "sex_male": 1,
  "smoker_yes": 0,
  "region_northwest": 0,
  "region_southeast": 1,
  "region_southwest": 0
}
'
```

Request URL

```
http://localhost:8000/predict
```

Server response

Code	Details
200	Response body

```
{
  "predicted_charges": 6859.69629231193
}
```

Download

Response headers

```
content-length: 38
content-type: application/json
date: Sun,20 Oct 2024 20:09:56 GMT
server: uvicorn
```

Responses

> The predicted charges for this particular input is about 6,859 dollars.
