medical_costs_post

September 29, 2024

1 Medical Insurance Cost prediction

This exercise is about performing some of the steps described in the notebook for the California Housing Data on another dataset for Medical Insurance Cost prediction.

2 Get the Data

[1]:
```
import pandas as pd

medical = pd.read_csv("https://bit.ly/44evDuW")
```

3 Take a Quick Look at the Data Structure

[2]:
```
# display the first 5 rows of the dataset by calling the head() function on␣
 ↪medical
medical.head()
```

[2]:

	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

Each row represents one patient. There are 7 attributes.

The `info()` method is useful to get a quick description of the data, in particular the total number of rows, each attribute's type, and the number of non-null values:

[3]:
```
# get the number of rows, columns, and data types by using the info() method
medical.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
```

[4]:
```
# show the number of patients in each region by using the value_counts() method␣
↪on the "region" column
medical.value_counts("region")
```

[4]:
```
region
southeast    364
northwest    325
southwest    325
northeast    324
Name: count, dtype: int64
```

Let's look at the other fields. The describe() method shows a summary of the numerical attributes.

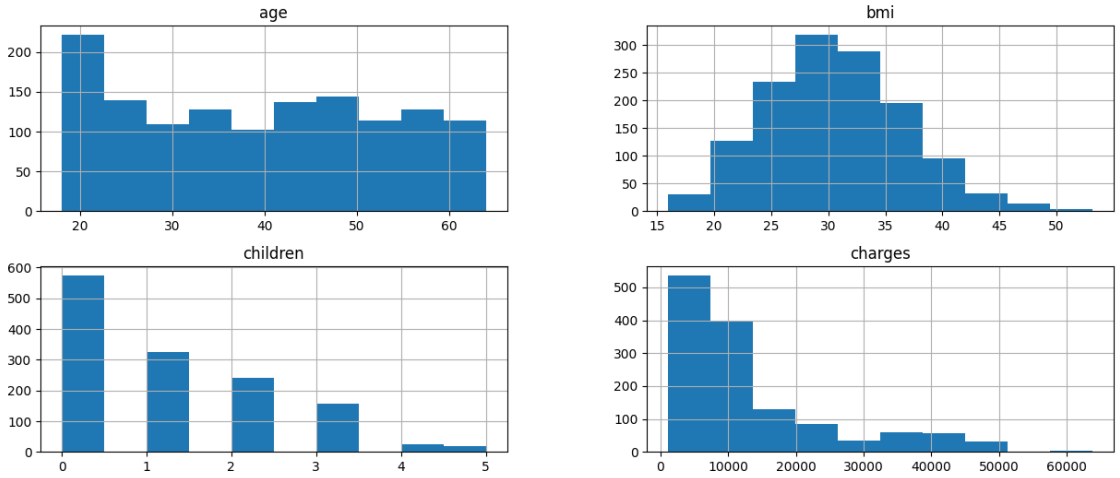
[5]:
```
# show descriptive statistics for the dataset by calling the describe() method␣
↪on medical
medical.describe()
```

[5]:

	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

[6]:
```
# show histograms for the numerical columns by using the hist() method on␣
↪medical
medical.hist(figsize=(15, 6))
```

[6]:
```
array([[<Axes: title={'center': 'age'}>, <Axes: title={'center': 'bmi'}>],
       [<Axes: title={'center': 'children'}>,
        <Axes: title={'center': 'charges'}>]], dtype=object)
```



Briefly write here what you observe from these histograms.

3.1 Create a Test Set

[7]:
```
# use train_test_split() to split the data into training and test sets
from sklearn.model_selection import train_test_split
train_set, test_set = train_test_split(medical, test_size=0.2, random_state=42)
```

4 Explore and Visualize the Data to Gain Insights

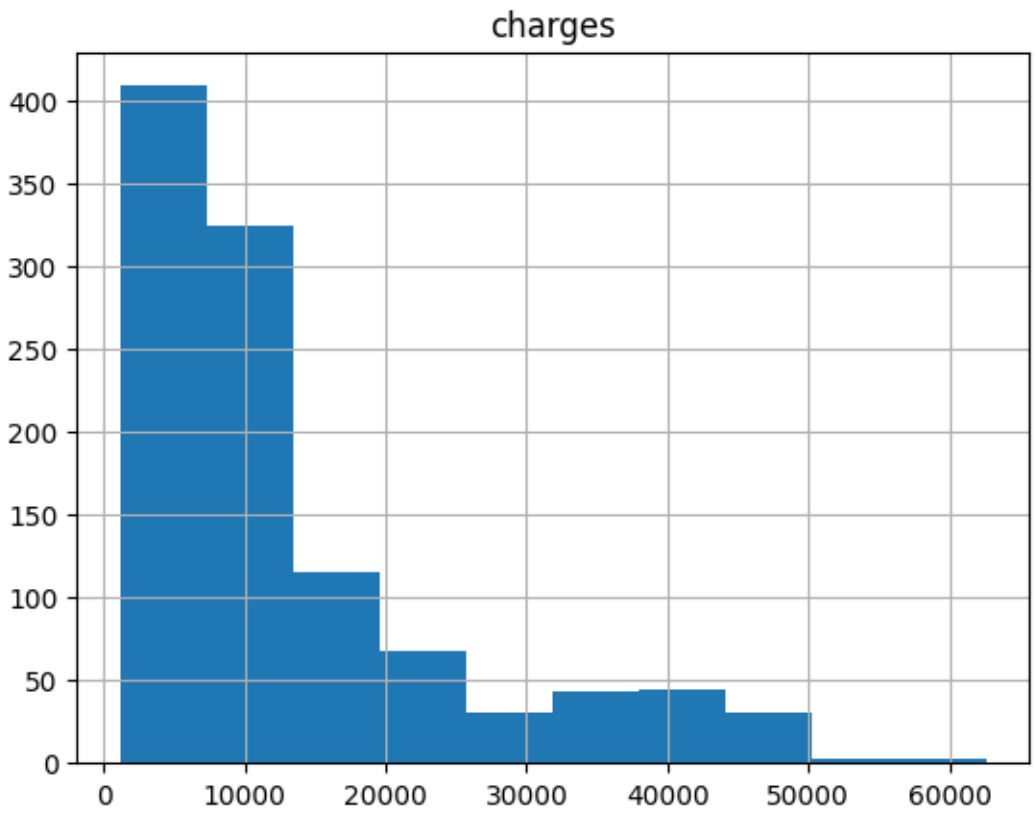
So far you have only taken a quick glance at the data to get a general understanding of the kind of data you are manipulating. Now the goal is to go into a little more depth.

First, make sure you have put the test set aside and you are only exploring the training set.

[8]:
```
# make a copy of the train set and save it to a variable called medical
medical = train_set
```

[9]:
```
# build a histogram of the charges column
medical.hist("charges")
```

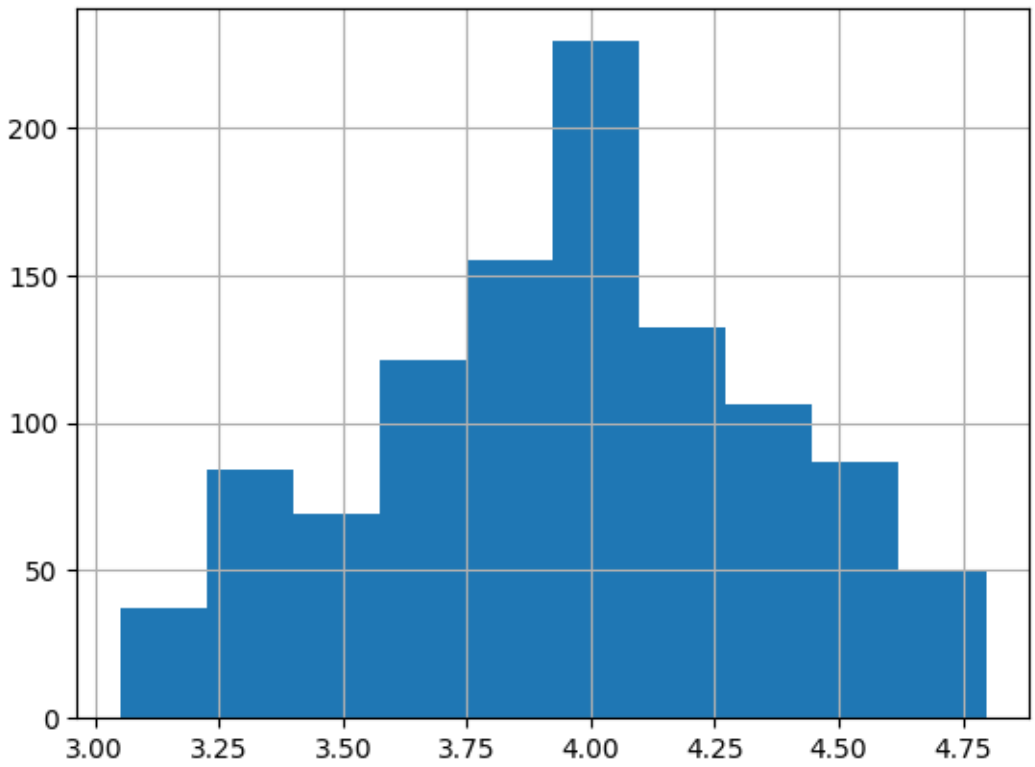
[9]:
```
array([[<Axes: title={'center': 'charges'}>]], dtype=object)
```



This distribution is right-skewed. To make it closer to normal we can apply natural log

[10]:
```
# apply a log transformation to the charges column using the np.log10() function
# build a histogram of the transformed column
import numpy as np
from sklearn.preprocessing import FunctionTransformer

log_transform = FunctionTransformer(np.log10)
log_charge = log_transform.fit_transform(medical["charges"])
log_charge.hist()
```

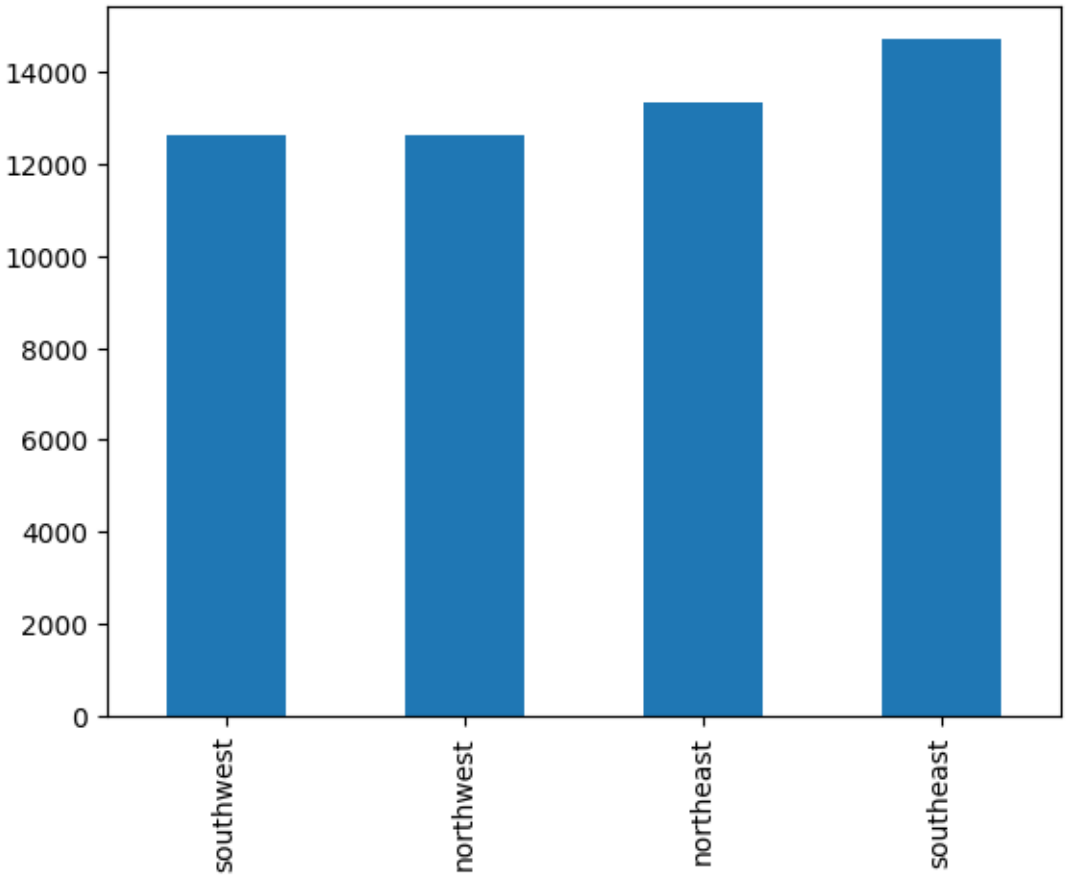
[10]: <Axes: >



Now let's look at the mean charges by region

[11]:
```
# compute the average insurance cost for each region
# sort the charges_by_region Series from the lowest to highest cost
# plot the sorted Series using the plot.bar() method
nw=medical.loc[medical["region"]=="northwest"]
ne=medical.loc[medical["region"]=="northeast"]
sw=medical.loc[medical["region"]=="southwest"]
se=medical.loc[medical["region"]=="southeast"]

dict = {"northwest": nw["charges"].mean(),
        "northeast": ne["charges"].mean(),
        "southwest": sw["charges"].mean(),
        "southeast": se["charges"].mean()}
charges_by_region = pd.Series(data=dict)
charges_by_region = charges_by_region.sort_values()

charges_by_region.plot.bar()
```

[11]: <Axes: >



Overall the highest medical charges are in the Southeast and the lowest are in the Southwest. Taking into account certain factors (sex, smoking, having children) let's see how it changes by region.

Now, create three grouped barcharts for average charges by region grouped by sex, smoking, and number of children.

4.0.1 How to create grouped barcharts?

Creating grouped bar charts with Seaborn is a bit more intuitive compared to Matplotlib. You can use the catplot function with kind='bar' to create grouped bar charts. Here is an example on the tips datasets that comes with Seaborn. The tips dataset contains information about the total bill and tip amount for different meals, along with additional information such as the sex of the individual paying for the meal, whether they are a smoker, the day and time of the meal, and the size of the party.

We will create a grouped bar chart showing the average total bill for each day, grouped by whether the meal took place at lunch or dinner.

```
import seaborn as sns
import matplotlib.pyplot as plt
```

```
# Load the 'tips' dataset
tips = sns.load_dataset("tips")

# Create a grouped bar chart
sns.catplot(data=tips, x="day", y="total_bill", hue="time", kind="bar")

plt.show()
```

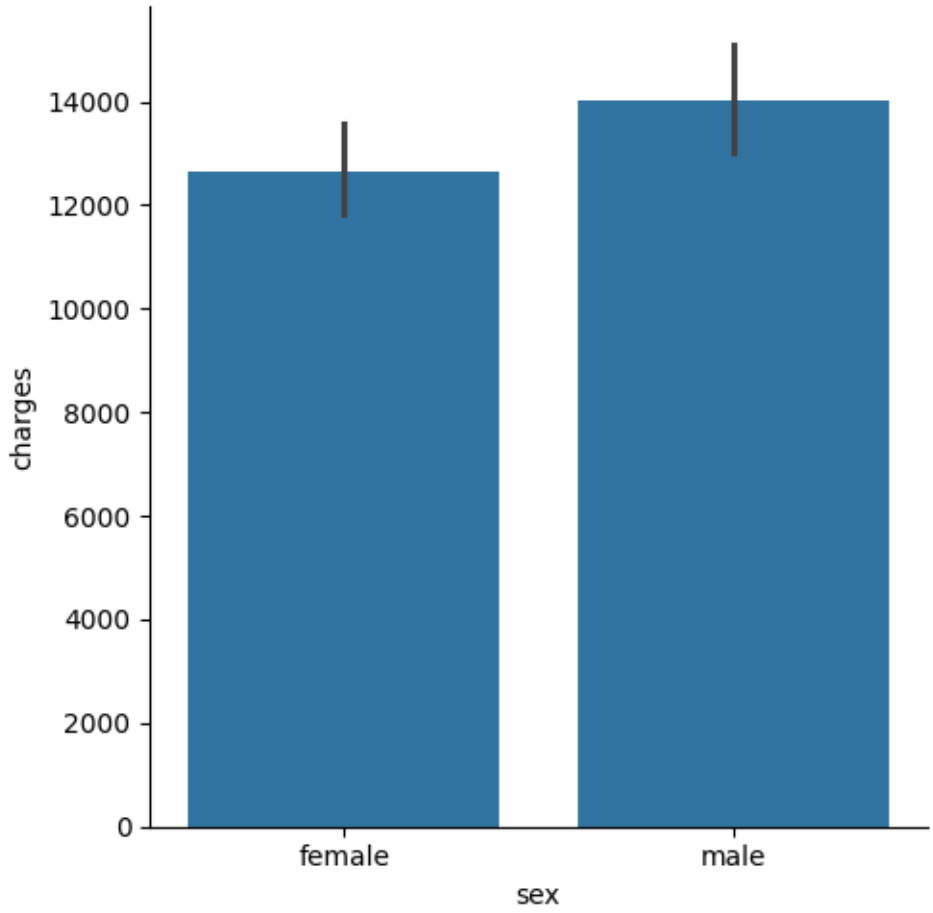
In this plot, the height of the bars represents the average total bill for meals on each day, with separate bars for lunch and dinner.

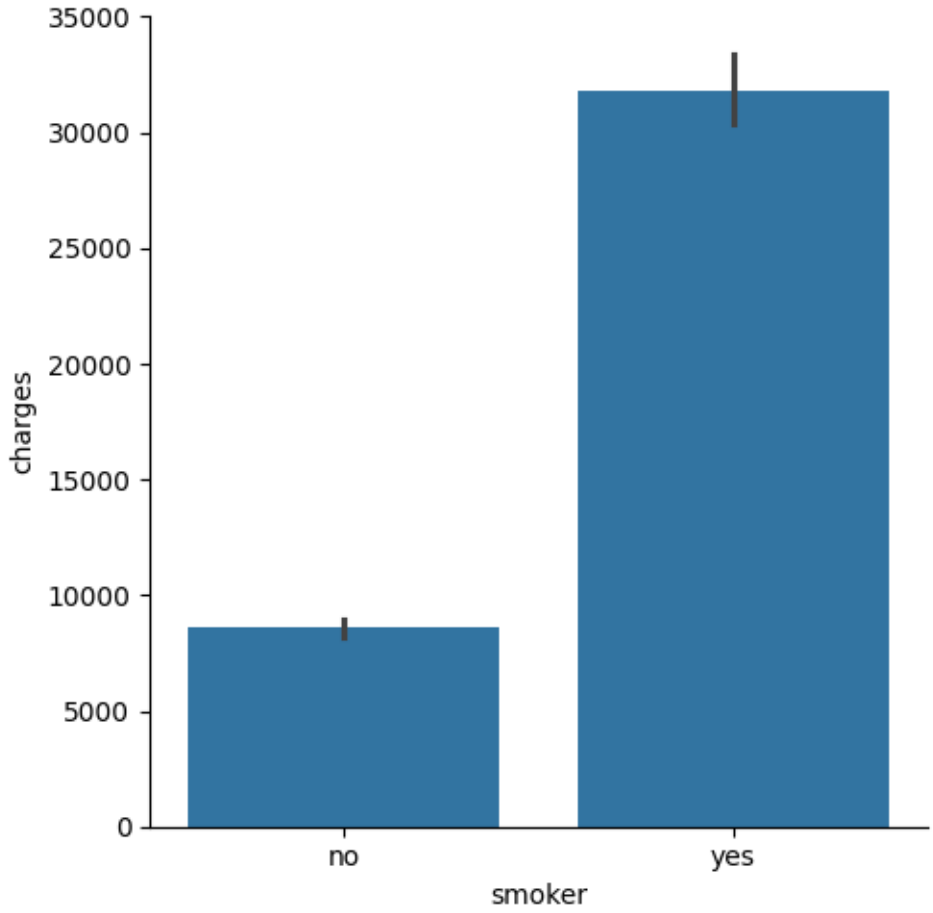
The `catplot` function is a flexible function that can create a variety of different plot types. By setting `kind='bar'`, we specify that we want a bar chart. The x and y arguments specify the data for the x and y axes, and the `hue` argument specifies a third variable that is used to group the data.

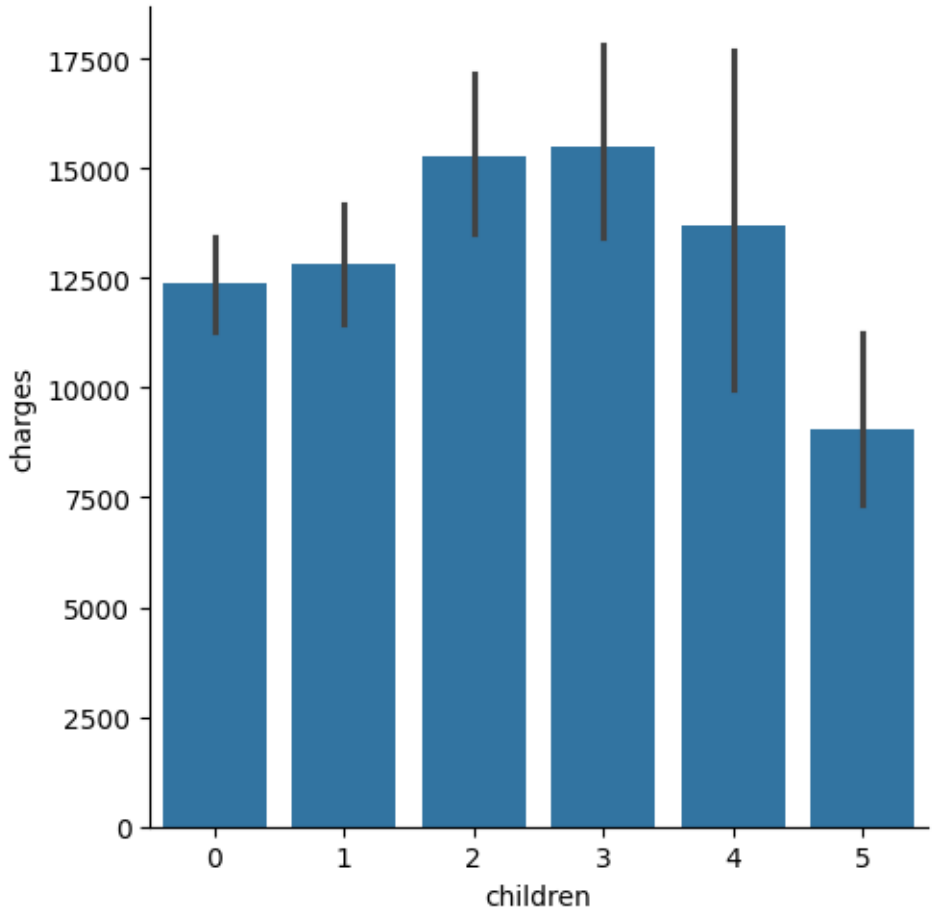
In the context of Seaborn and many other statistical visualization libraries, error bars commonly represent one standard deviation or standard error of the mean.

[12]:
```
# plot grouped bar charts of region and insurance costs hue by sex, smoker, and␣
 ↪number of children (three separate charts)
# use the catplot() function to create the bar charts
# set the kind parameter to "bar" and the data parameter to medical
import seaborn as sns
import matplotlib as plt
sns.catplot(data=medical,
            kind="bar",
            x="sex",
            y="charges",)
sns.catplot(data=medical,
            kind="bar",
            x="smoker",
            y="charges",)
sns.catplot(data=medical,
            kind="bar",
            x="children",
            y="charges",)
```

[12]:
```
<seaborn.axisgrid.FacetGrid at 0x730ef8e559d0>
```


14000
12000
10000
8000
6000
4000
2000
0
charges
female
male
sex


35000
30000
25000
20000
15000
10000
5000
0
charges
no
yes
smoker



What do you observe? Briefly write what you observe from the charts.

Student answer

Sex does not appear to create a large difference, in this set males are charged more but more analysis is warranted to see if this is the expected result Smokers are charged far more than non-smokers. There appears to be a large drop off in charges for people with 4-5 children, with 3 children being the most expensive. I wonder why that's the case

---

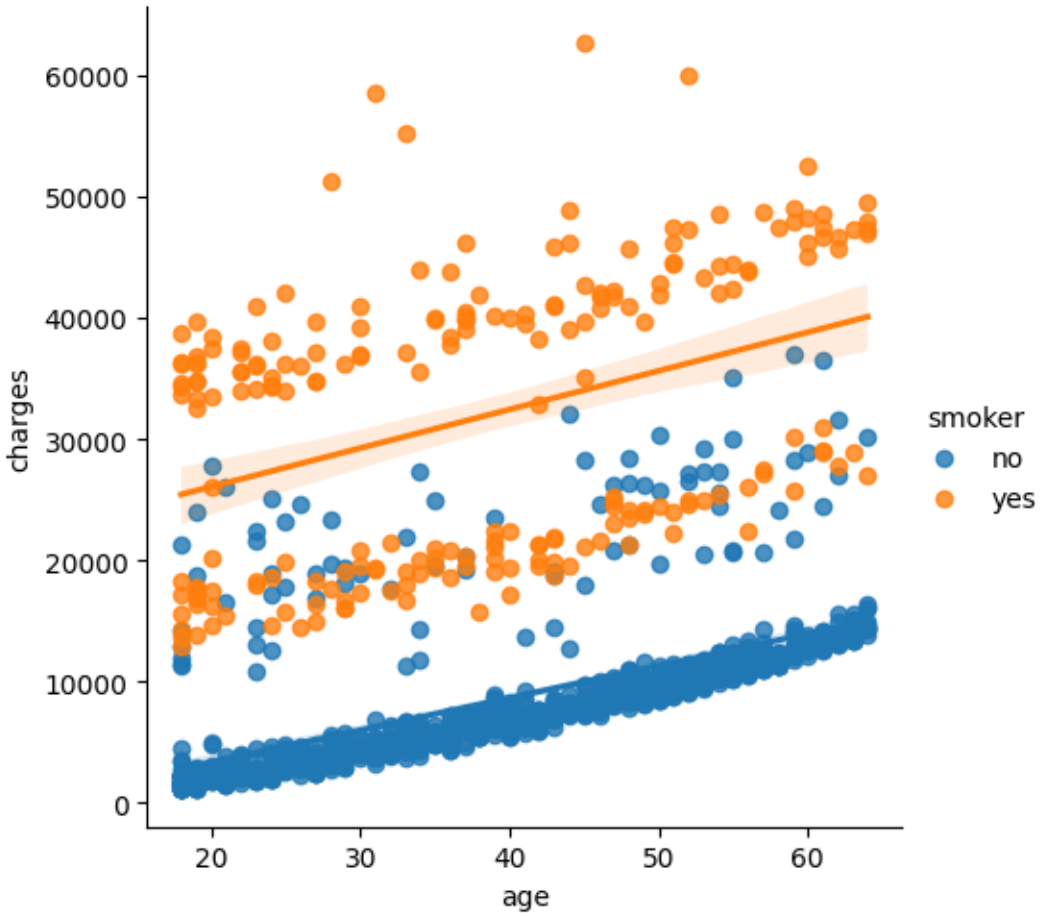
Now let's analyze the medical charges by age, bmi and children according to the smoking factor.

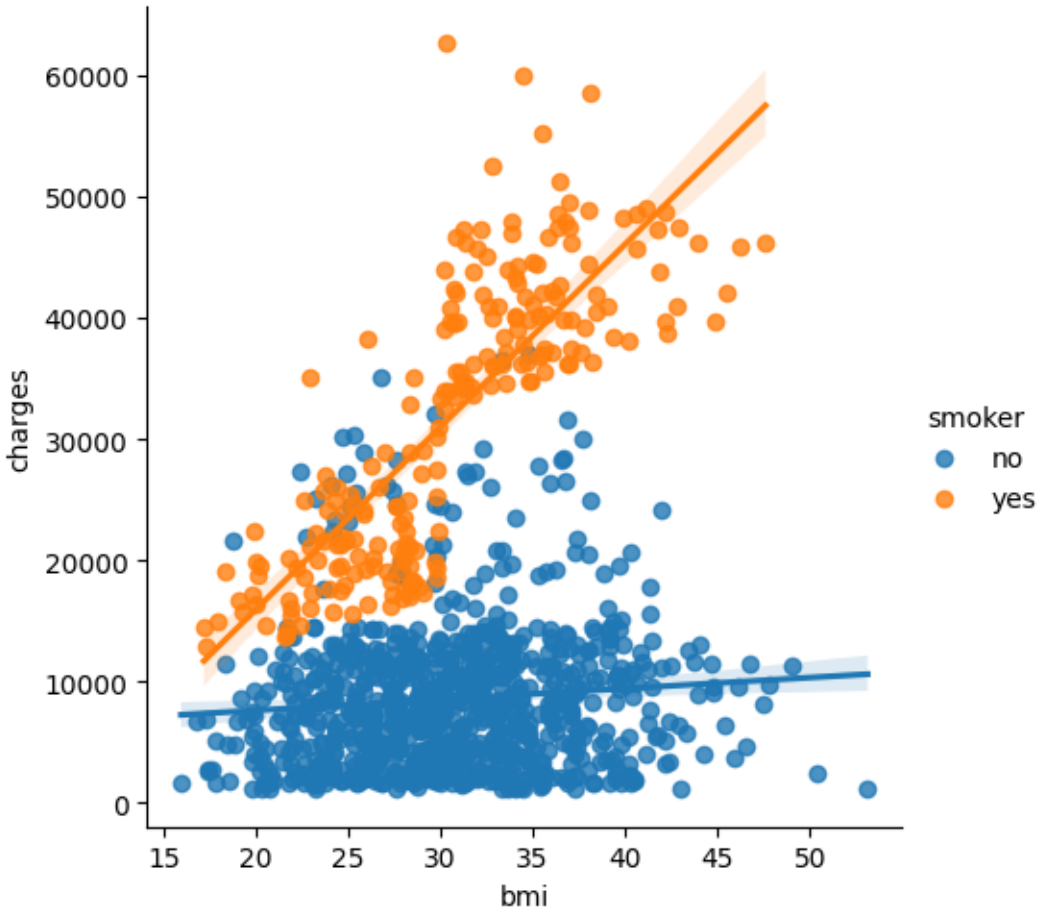
[13]:
```
# using the lmplot() function of seaborn, build a scatter plot of age and␣
↪insurance costs, hue by smoker
# build a second scatter plot of bmi and insurance costs, hue by smoker
# build a third scatter plot of children and insurance costs, hue by smoker
medical.info()
sns.lmplot(data=medical,
```

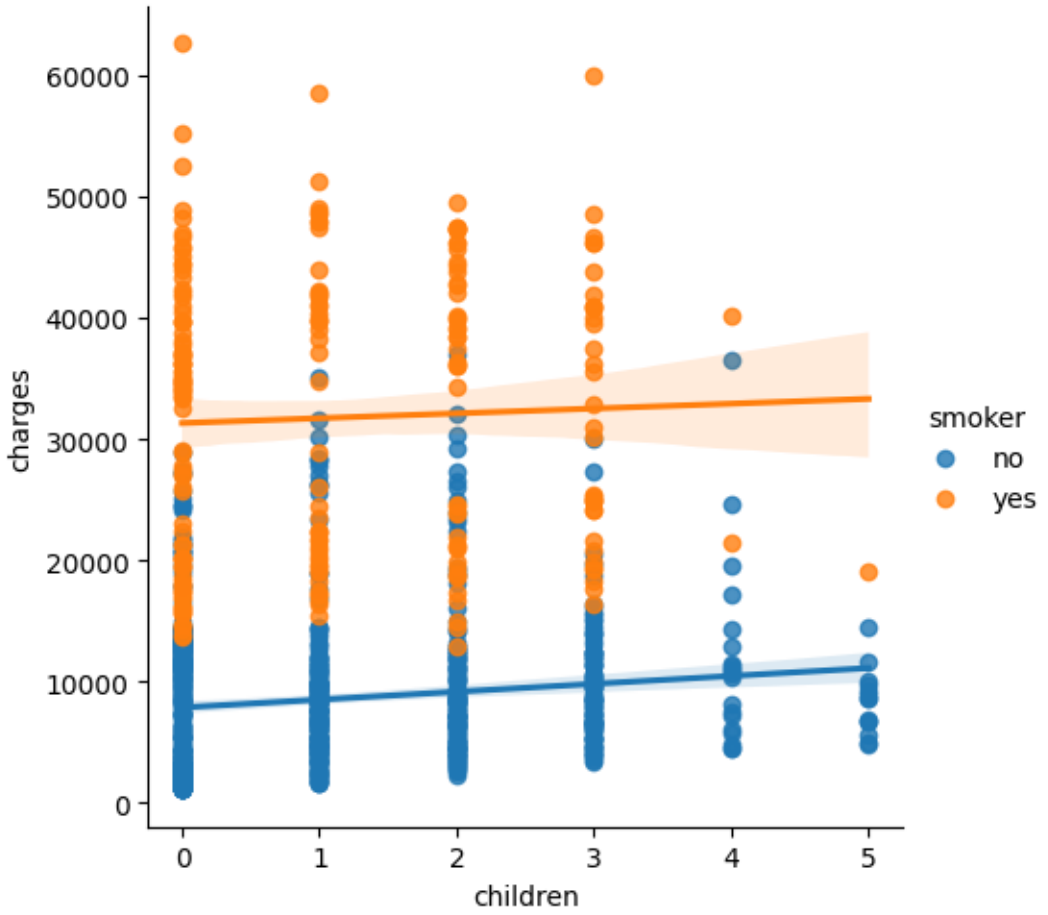
```
            x="age",
            y="charges",
            hue="smoker")

sns.lmplot(data=medical,
            x="bmi",
            y="charges",
            hue="smoker")

sns.lmplot(data=medical,
            x="children",
            y="charges",
            hue="smoker")
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1070 entries, 560 to 1126
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1070 non-null   int64
 1   sex       1070 non-null   object
 2   bmi       1070 non-null   float64
 3   children  1070 non-null   int64
 4   smoker    1070 non-null   object
 5   region    1070 non-null   object
 6   charges   1070 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 66.9+ KB
```

[13]: <seaborn.axisgrid.FacetGrid at 0x730f2bfae840>


60000
50000
40000
30000
20000
10000
0
charges
20
30
40
50
60
age
smoker
no
yes


60000
50000
40000
30000
20000
10000
0
charges
15
20
25
30
35
40
45
50
bmi
smoker
no
yes



Describe in a one-liner what you obeserve from the charts.

Smokers consistently pay more, insurance costs more as you age or as BMI increases, and in this dataset, parents to 4 or 5 children don't smoke as much

4.0.2 Look for Correlations

[14]:
```
# compute pairwise correlation of columns using the corr() method

#first, need to separate out numerical columns
medical_num = medical.select_dtypes(include=[np.number])
medical_num.corr()
```

[14]:

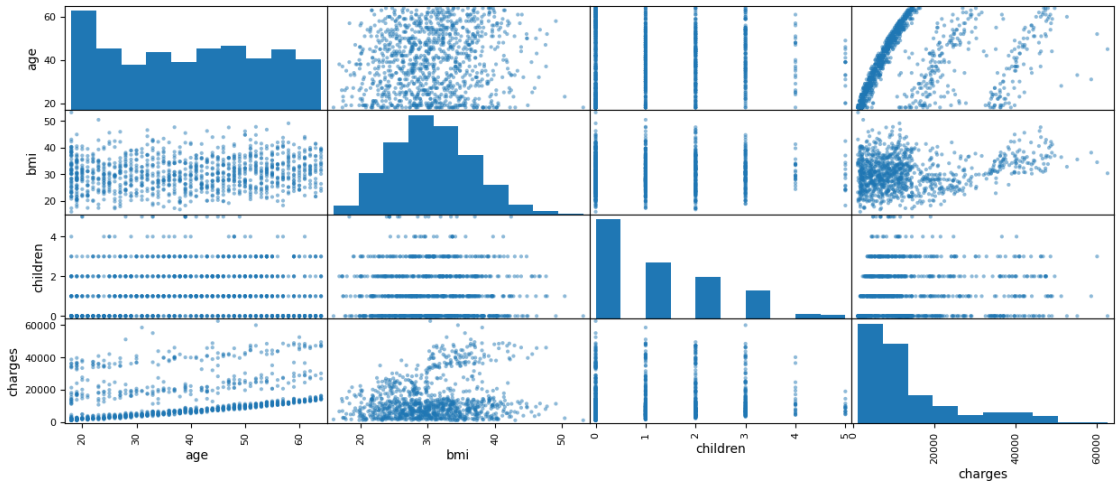
	age	bmi	children	charges
age	1.000000	0.118274	0.060999	0.281721
bmi	0.118274	1.000000	-0.005040	0.197316
children	0.060999	-0.005040	1.000000	0.071885
charges	0.281721	0.197316	0.071885	1.000000

The correlation coefficient ranges from –1 to 1. When it is close to 1, it means that there is a strong positive correlation. Finally, coefficients close to 0 mean that there is no linear correlation.

Another way to check for correlation between attributes is to use the Pandas `scatter_matrix()` function, which plots every numerical attribute against every other numerical attribute.

[15]:
```
# plot correlation matrix using scatter_matrix() function from pandas.plotting
pd.plotting.scatter_matrix(frame=medical, figsize=(15, 6))
```

[15]:
```
array([[<Axes: xlabel='age', ylabel='age'>,
        <Axes: xlabel='bmi', ylabel='age'>,
        <Axes: xlabel='children', ylabel='age'>,
        <Axes: xlabel='charges', ylabel='age'>],
       [<Axes: xlabel='age', ylabel='bmi'>,
        <Axes: xlabel='bmi', ylabel='bmi'>,
        <Axes: xlabel='children', ylabel='bmi'>,
        <Axes: xlabel='charges', ylabel='bmi'>],
       [<Axes: xlabel='age', ylabel='children'>,
        <Axes: xlabel='bmi', ylabel='children'>,
        <Axes: xlabel='children', ylabel='children'>,
        <Axes: xlabel='charges', ylabel='children'>],
       [<Axes: xlabel='age', ylabel='charges'>,
        <Axes: xlabel='bmi', ylabel='charges'>,
        <Axes: xlabel='children', ylabel='charges'>,
        <Axes: xlabel='charges', ylabel='charges'>]], dtype=object)
```

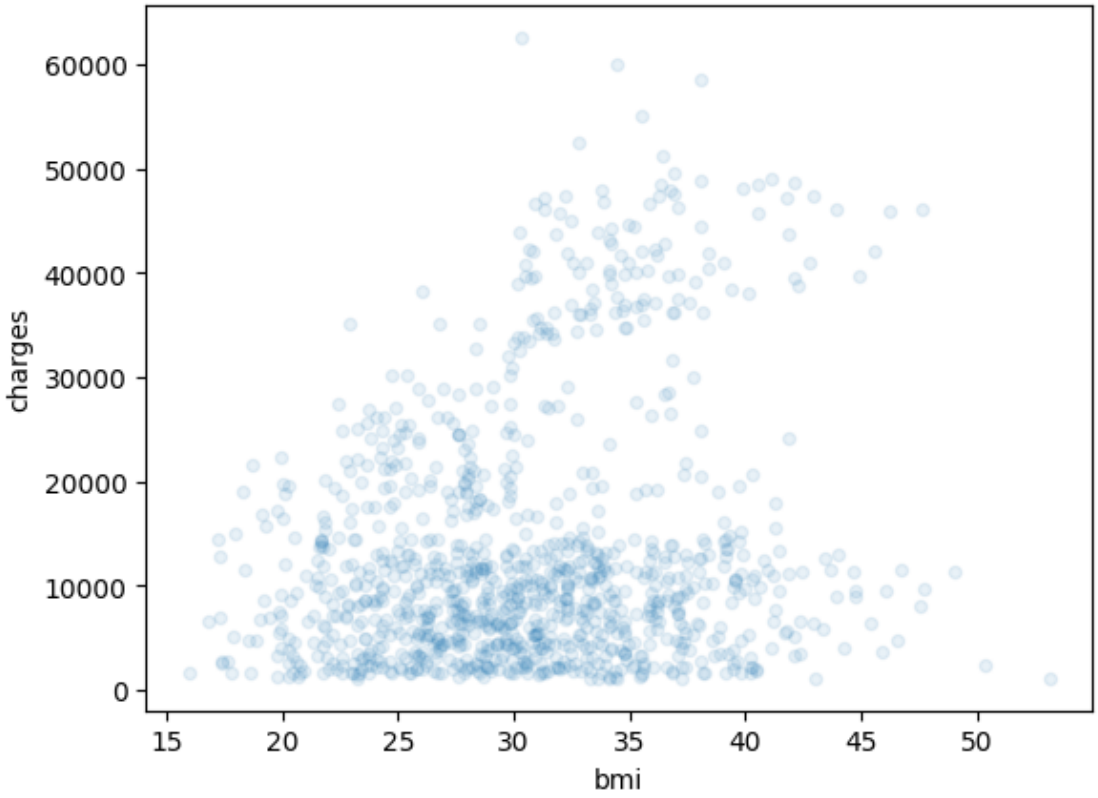


The main diagonal would be full of straight lines if Pandas plotted each variable against itself, which would not be very useful. So instead, the Pandas displays a histogram of each attribute.

Looking at the correlation scatterplots, it seems like the most promising attribute to predict the charge value is bmi, so let's zoom in on their scatterplot.

[16]:
```
# plot a scatter plot of bmi vs. insurance costs using the medical.plot()␣
↪method, use the alpha parameter to set the opacity of the points to 0.1
medical.plot(kind="scatter",
             x="bmi",
             y="charges",
             alpha=0.1)
```

[16]: <Axes: xlabel='bmi', ylabel='charges'>



The correlation is somewhat visible; you can clearly see the upward trend.

5 Prepare the data for ML

[17]:
```
# drop the charges column from the train_set and save the resulting dataset to␣
 ↪a variable called `medical`
# create a copy of the train_set labels and save it to a variable called␣
 ↪`medical_labels`
# replace None with the correct code

medical = train_set.drop("charges", axis=1)
medical_labels = train_set["charges"].copy()
```

6 Transformation Pipelines

As you can see, there are many data transformation steps that need to be executed in the right order. Fortunately, Scikit-Learn provides the Pipeline class to help with such sequences of transformations.

[18]:
```
# uncomment the following code to create a pipeline for preprocessing the data

from sklearn.compose import ColumnTransformer
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import OneHotEncoder
from sklearn.pipeline import make_pipeline

num_attribs = ["age", "bmi", "children"]
cat_attribs = ["sex", "smoker", "region"]

num_pipeline = make_pipeline(
    SimpleImputer(strategy="median"),
    StandardScaler())

cat_pipeline = make_pipeline(
    SimpleImputer(strategy="most_frequent"),
    OneHotEncoder(handle_unknown="ignore"))

preprocessing = ColumnTransformer([
    ("num", num_pipeline, num_attribs),
    ("cat", cat_pipeline, cat_attribs)])

medical_prepared = preprocessing.fit_transform(medical)

print(medical_prepared.shape)
print(preprocessing.get_feature_names_out())
```

```
(1070, 11)
['num__age' 'num__bmi' 'num__children' 'cat__sex_female' 'cat__sex_male'
 'cat__smoker_no' 'cat__smoker_yes' 'cat__region_northeast'
 'cat__region_northwest' 'cat__region_southeast' 'cat__region_southwest']
```

7 Select and Train a Model

At last! You framed the problem, you got the data and explored it, you sampled a training set and a test set, and you wrote a preprocessing pipeline to automatically clean up and prepare your data for machine learning algorithms. You are now ready to select and train a machine learning model.

7.1 Train and Evaluate on the Training Set

The good news is that thanks to all these previous steps, things are now going to be easy! You decide to train a very basic linear regression model to get started:

[19]:
```
# create a pipeline for preprocessing the data and fitting a linear regression␣
 ↪model
# lin_reg = ...
from sklearn.linear_model import LinearRegression
lin_reg = make_pipeline(preprocessing, LinearRegression())

# housing_labels is the column we want to predict
# uncomment the following line to fit the model

lin_reg.fit(medical, medical_labels)
```

[19]:
```
Pipeline(steps=[('columntransformer',
                 ColumnTransformer(transformers=[('num',
Pipeline(steps=[('simpleimputer',
SimpleImputer(strategy='median')),
('standardscaler',
StandardScaler())]),
                                                 ['age', 'bmi', 'children']),
                                                ('cat',
Pipeline(steps=[('simpleimputer',
SimpleImputer(strategy='most_frequent')),
('onehotencoder',
OneHotEncoder(handle_unknown='ignore'))]),
                                                 ['sex', 'smoker',
                                                  'region'])])),
                ('linearregression', LinearRegression())])
```

Try out the model on the training set, look at the first five predictions and compare them to the labels:

[20]:
```
# uncomment the following line to make predictions

medical_predictions = lin_reg.predict(medical)
medical_predictions
```

[20]:
```
array([ 7094.54007011,  8344.72998713,  9153.77419778, …,
       11441.08519155, 37314.37460682, 11453.12102783])
```

[21]:
```
# uncomment the following lines to compute the RMSE

from sklearn.metrics import mean_squared_error
from sklearn.metrics import root_mean_squared_error

# lin_rmse = mean_squared_error(medical_labels, medical_predictions,␣
 ↪squared=False) DEPRECIATED
lin_rmse = root_mean_squared_error(medical_labels, medical_predictions)
lin_rmse
```

[21]:
```
np.float64(6105.545160099848)
```

Now try `DecisionTreeRegressor`, as this is a fairly powerful model capable of finding complex nonlinear relationships in the data (decision trees are covered later in the course):

[22]:
```
# use DecisionTreeRegressor to train the model
# use the make_pipeline() function to create a pipeline for preprocessing and␣
 ↪model training
# use the preprocessing object you created earlier
# make predictions on the training set and compute the RMSE

from sklearn.tree import DecisionTreeRegressor

tree_reg = make_pipeline(preprocessing, DecisionTreeRegressor())
tree_reg.fit(medical, medical_labels)

tree_predictions = tree_reg.predict(medical)
tree_rmse = root_mean_squared_error(medical_labels, tree_predictions)
tree_rmse
```

[22]:
```
np.float64(494.20598375812835)
```

8 Better Evaluation Using Cross-Validation

The following code randomly splits the training set into 10 nonoverlapping subsets called folds, then it trains and evaluates the decision tree model 10 times, picking a different fold for evaluation every time and using the other 9 folds for training. The result is an array containing the 10 evaluation scores:

[23]:
```
# uncomment the following lines to train the model and make predictions

from sklearn.model_selection import cross_val_score

tree_rmses = -cross_val_score(tree_reg,
                              medical,medical_labels,
                              scoring="neg_root_mean_squared_error",
                              cv=10)
```

```
tree_rmses
```

[23]:
```
array([6543.90885675, 6520.63610249, 6568.00216288, 6876.99024608,
       7127.34862151, 6716.49482643, 7358.25706427, 6838.54695789,
       6829.74874254, 5179.08566961])
```

Warning. Scikit-Learn's cross-validation features expect a utility function (greater is better) rather than a cost function (lower is better), so the scoring function is actually the opposite of the RMSE. It's a negative value, so you need to switch the sign of the output to get the RMSE scores.

[24]:
```
# uncomment the following line to compute the mean of the RMSEs

np.mean(tree_rmses)
```

[24]:
```
np.float64(6655.9019250459005)
```

Let's try one last model now: the RandomForestRegressor. As you will see later in the course, random forests work by training many decision trees on random subsets of the features, then averaging out their predictions.

[25]:
```
# use RandomForestRegressor to train the model
# use the make_pipeline() function to create a pipeline for preprocessing and␣
 ↪model training
# use the preprocessing object you created earlier
# make predictions on the training set and compute the RMSEs using␣
 ↪cross-validation
# compute the mean of the RMSEs
from sklearn.ensemble import RandomForestRegressor

forest_reg = make_pipeline(preprocessing, RandomForestRegressor())
forest_reg.fit(medical, medical_labels)

forest_predictions = forest_reg.predict(medical)
forest_rmse = root_mean_squared_error(medical_labels, forest_predictions)
print(forest_rmse)

forest_rmses = -cross_val_score(forest_reg,
                                medical,medical_labels,
                                scoring="neg_root_mean_squared_error",
                                cv=10)
print(forest_rmses)
print(np.mean(forest_rmses))
```

```
1895.3282588748175
[4743.50409232 5214.03474052 3855.10375506 4615.56242803 5237.26115273
 4969.98503617 5632.57668005 5447.42784546 5184.53591719 4703.68083525]
4960.36724827794
```

9 Fine-Tune Your Model

Let's assume that you now have a shortlist of promising models. You now need to fine-tune them.

9.1 Randomized Search for Good Hyperparameters

[26]:
```
# uncomment the following lines to search for the best hyperparameters

from sklearn.model_selection import RandomizedSearchCV
from scipy.stats import randint
from sklearn.pipeline import Pipeline

full_pipeline = Pipeline([("preprocessing", preprocessing),
                          ("random_forest",␣
 ↪RandomForestRegressor(random_state=42)),
                          ])

param_distribs = {'random_forest__max_features': randint(low=2,high=20)}

rnd_search = RandomizedSearchCV(full_pipeline,
                                param_distributions=param_distribs,
                                n_iter=10,
                                cv=3,
                                scoring='neg_root_mean_squared_error',
                                random_state=42)

rnd_search.fit(medical, medical_labels)
```

[26]:
```
RandomizedSearchCV(cv=3,
                   estimator=Pipeline(steps=[('preprocessing',
ColumnTransformer(transformers=[('num',
Pipeline(steps=[('simpleimputer',
                 SimpleImputer(strategy='median')),
                ('standardscaler',
                 StandardScaler())]),
['age',
'bmi',
'children']),
('cat',
Pipeline(steps=[('simpleimputer',
                 SimpleImputer(strategy='most_frequent')),
                ('onehotencoder',
                 OneHotEncoder(handle_unknown='ignore'))]),
['sex',
'smoker',
'region'])])),
                                                ('random_forest',
```

```
RandomForestRegressor(random_state=42))]),
                   param_distributions={'random_forest__max_features':
<scipy.stats._distn_infrastructure.rv_discrete_frozen object at
0x730ef60f8050>},
                   random_state=42, scoring='neg_root_mean_squared_error')
```

[27]:
```
# uncomment the following lines to print the best search scores

rn_res = pd.DataFrame(rnd_search.cv_results_)
rn_res.sort_values(by="mean_test_score", ascending=False, inplace=True)
rn_res.head(5)["mean_test_score"]
```

[27]:
```
7   -4810.698561
9   -4861.373750
4   -4912.976290
0   -4912.976290
8   -4976.129581
Name: mean_test_score, dtype: float64
```

[28]:
```
# uncomment the following lines to print the feature importances

final_model = rnd_search.best_estimator_ # includes preprocessing
feature_importances = final_model["random_forest"].feature_importances_
feature_importances
```

[28]:
```
array([0.14032922, 0.18831699, 0.0230048 , 0.00493845, 0.00454404,
       0.28737838, 0.33010198, 0.00580561, 0.00521445, 0.00654471,
       0.00382136])
```

[29]:
```
# uncomment the following line to print the feature importances with the␣
 ↪feature names

sorted(zip(feature_importances, final_model["preprocessing"].
 ↪get_feature_names_out()),reverse=True)
```

[29]:
```
[(np.float64(0.3301019845385152), 'cat__smoker_yes'),
 (np.float64(0.2873783836848969), 'cat__smoker_no'),
 (np.float64(0.18831699212058217), 'num__bmi'),
 (np.float64(0.1403292169242826), 'num__age'),
 (np.float64(0.023004795671815754), 'num__children'),
 (np.float64(0.006544713122635331), 'cat__region_southeast'),
 (np.float64(0.005805609883995695), 'cat__region_northeast'),
 (np.float64(0.005214451279002379 5), 'cat__region_northwest'),
 (np.float64(0.0049384524581677835), 'cat__sex_female'),
 (np.float64(0.004544038586538502), 'cat__sex_male'),
 (np.float64(0.0038213617295676603), 'cat__region_southwest')]
```

[30]:
```
# now that you have a final model, evaluate it on the test set (find rmse)
final_predictions = final_model.predict(medical)
final_rmse = root_mean_squared_error(medical_labels, final_predictions)
print(final_rmse)
```

```
1857.6573579690826
```