Problem 4

Objectives

1. Calculate and interpret the correlation matrix to understand relationships among features.
2. Create a scatterplot matrix to visualize relationships among features. Explain the insights they can gain from these visualizations.
3. Perform data preprocessing and cleaning, which involves addressing missing values and handling categorical features, followed by conducting a train-test split of the data.
4. Implementing and training the linear regression model (apply Ridge and Lasso regression techniques) using appropriate Python libraries.
5. Evaluate the model's performance by calculating relevant metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared. Additionally, interpret the model's coefficients and discuss how various features impact predictions of medical expenses.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression, Ridge, Lasso
from sklearn.metrics import mean_absolute_error, mean_squared_error,
r2_score
from sklearn import preprocessing

path = '/content/drive/MyDrive/sem 7/ID5055/Assignment 3/Problem
4/insurance.csv'

data = pd.read_csv(path)
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   expenses  1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

There are intotal 1338 entries for 7 features and we have 3 'object' datatype features (sex, smoker, and region).

```
correlation = data.corr()
correlation
```

```
<ipython-input-417-521f87fcc686>:1: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a future version, it will default to False. Select only valid columns or specify the value of numeric_only to silence this warning.
  correlation = data.corr()
```

```
               age       bmi  children  expenses
age       1.000000  0.109341  0.042469  0.299008
bmi       0.109341  1.000000  0.012645  0.198576
children  0.042469  0.012645  1.000000  0.067998
expenses  0.299008  0.198576  0.067998  1.000000
```

From correlation matrix between numerical datasets it is clear that-

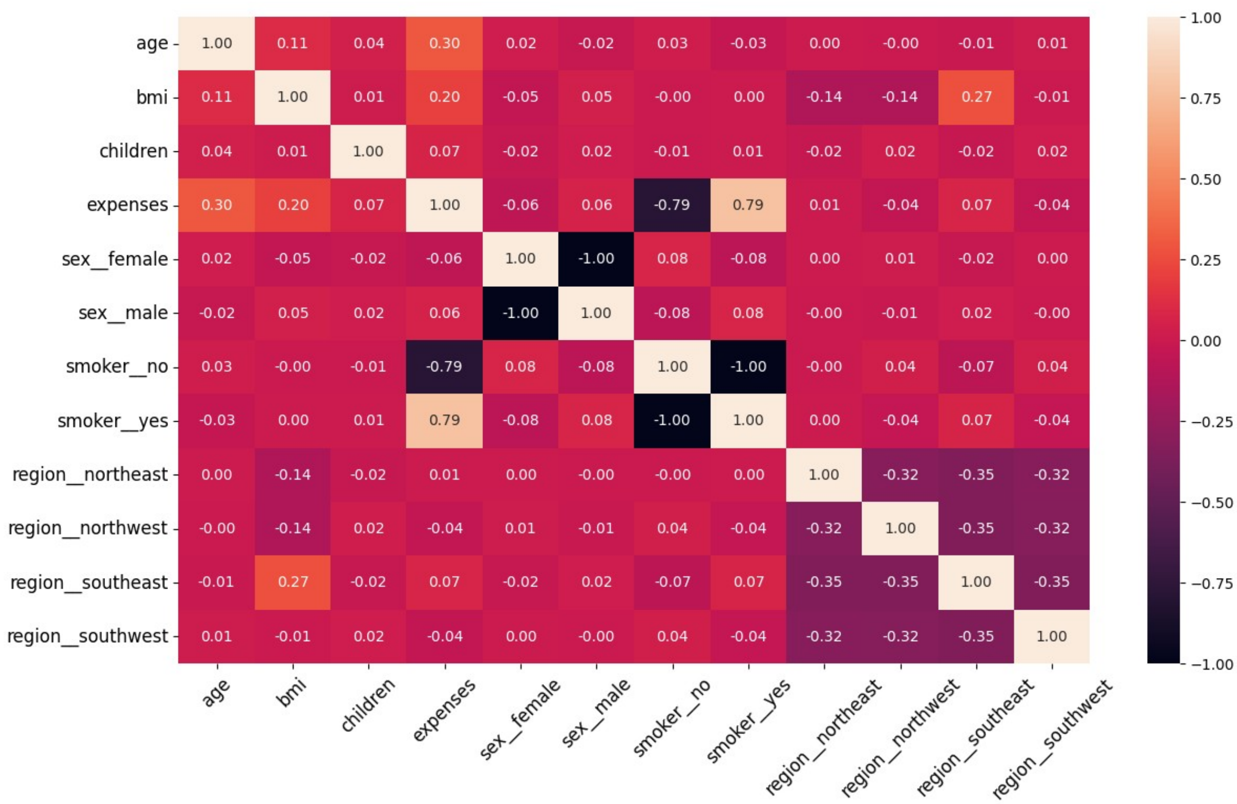
1. Age and bmi associate strongly with expenses and otherway round i.e. expenses associate with age and bmi.
2. The expenses does not associate that strongly with number of children.
3. BMI and age are correlated weakly.
4. We still have to check the categorical features to get a better idea.

```
df2 = data.copy()

sex_dummies = pd.get_dummies(df2['sex'], prefix = 'sex_')
df2.drop(['sex'], axis = 1, inplace = True)
df2 = pd.concat([df2, sex_dummies], axis = 1)

smoker_dummies = pd.get_dummies(df2['smoker'], prefix = 'smoker_')
df2.drop(['smoker'], axis = 1, inplace = True)
df2 = pd.concat([df2, smoker_dummies], axis = 1)

region_dummies = pd.get_dummies(df2['region'], prefix = 'region_')
df2.drop(['region'], axis = 1, inplace = True)
df2 = pd.concat([df2, region_dummies], axis = 1)

df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   age                1338 non-null   int64  
 1   bmi                1338 non-null   float64
 2   children           1338 non-null   int64  
 3   expenses           1338 non-null   float64
```

```
 4   sex__female        1338 non-null   uint8
 5   sex__male          1338 non-null   uint8
 6   smoker__no         1338 non-null   uint8
 7   smoker__yes        1338 non-null   uint8
 8   region__northeast  1338 non-null   uint8
 9   region__northwest  1338 non-null   uint8
 10  region__southeast  1338 non-null   uint8
 11  region__southwest  1338 non-null   uint8
dtypes: float64(2), int64(2), uint8(8)
memory usage: 52.4 KB
```

```
fig2, ax2 = plt.subplots(figsize=(14, 8))
corr_matrix_2 = df2.corr()

sns.heatmap(corr_matrix_2, annot=True, xticklabels=True,
yticklabels=True,
            annot_kws={"size": 10}, fmt=f'.{2}f', ax=ax2)
ax2.set_yticklabels(ax2.get_yticklabels(), rotation=0)
ax2.set_xticklabels(ax2.get_xticklabels(), rotation=45)
ax2.tick_params(axis='both', which='both', labelsize=12)

plt.show()
```



Observations and insights

1. Age vs. Expenses: Age has a positive correlation of approximately 0.30 with expenses. This suggests that as people get older, their medical expenses tend to increase. This correlation is moderately strong.

2. BMI vs. Expenses: BMI (Body Mass Index) also has a positive correlation with expenses, but it is weaker compared to age, with a correlation of approximately 0.20. This indicates that individuals with higher BMIs tend to have somewhat higher medical expenses.

3. Smoking Status vs. Expenses: Smoking status has a strong correlation with expenses. "smoker__yes" (indicating a smoker) has a positive correlation of approximately 0.79 with expenses, while "smoker__no" (indicating a non-smoker) has a negative correlation of approximately -0.79. This indicates that smokers tend to have significantly higher medical expenses compared to non-smokers.

4. Region vs. Expenses: The region where a person lives also has some correlation with expenses, although these correlations are relatively weak. None of the regional variables have a strong impact on medical expenses, but there are some variations.

5. Gender vs. Expenses: Gender has a relatively weak correlation with expenses. "sex__female" has a negative correlation of approximately -0.06, while "sex__male" has a positive correlation of approximately 0.06. This suggests that, on average, females may have slightly lower medical expenses than males in the dataset, although the effect is not very significant.

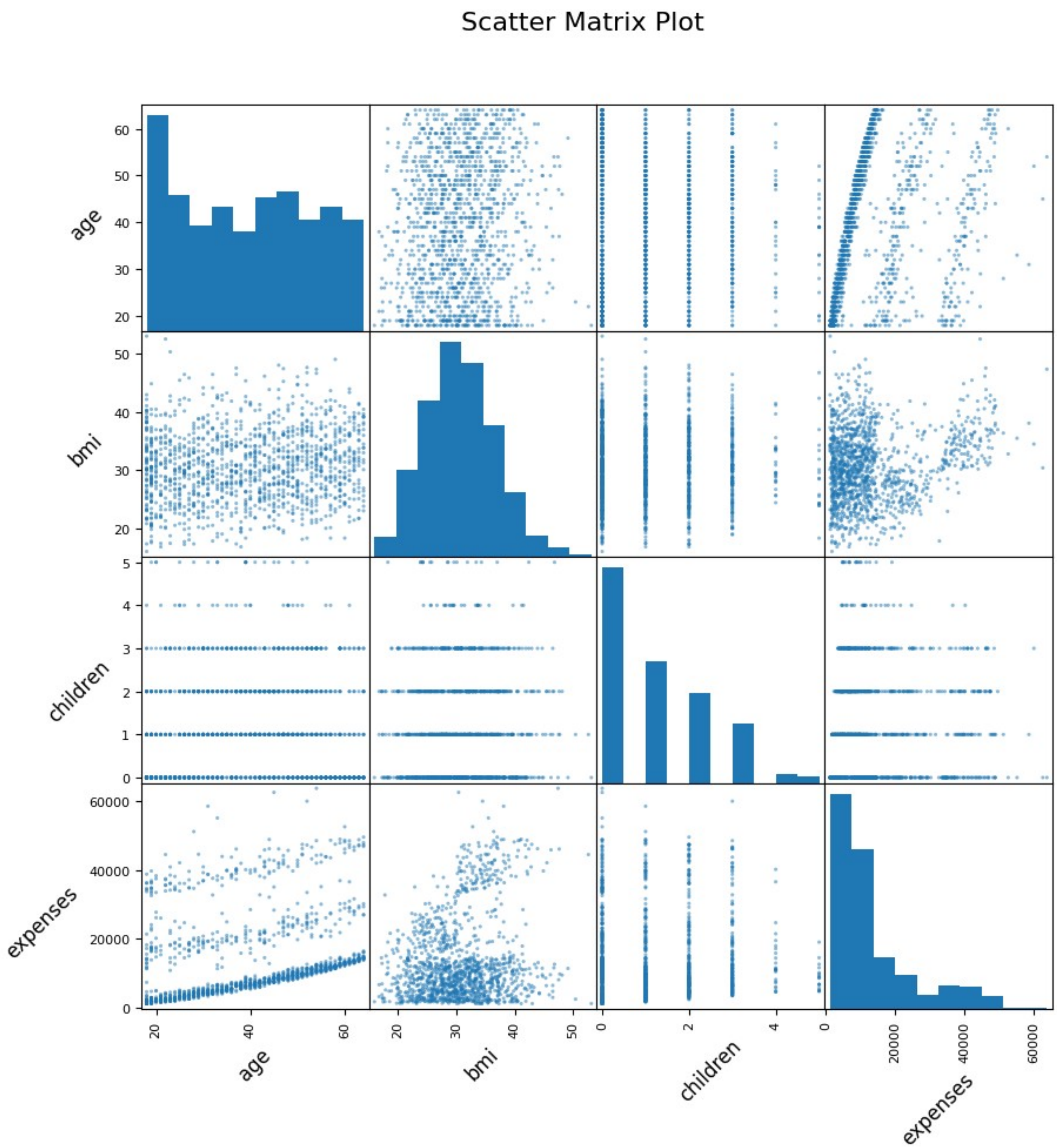
6. Number of Children vs. Expenses: The number of children a person has ("children" variable) has a relatively weak positive correlation of approximately 0.07 with expenses. This implies that individuals with more children may have slightly higher medical expenses, but the effect is not very strong.

```
df_plot = data.copy()
df_plot
```

	age	sex	bmi	children	smoker	region	expenses
0	19	female	27.9	0	yes	southwest	16884.92
1	18	male	33.8	1	no	southeast	1725.55
2	28	male	33.0	3	no	southeast	4449.46
3	33	male	22.7	0	no	northwest	21984.47
4	32	male	28.9	0	no	northwest	3866.86
...	...	...	...	...	...	...	...
1333	50	male	31.0	3	no	northwest	10600.55
1334	18	female	31.9	0	no	northeast	2205.98
1335	18	female	36.9	0	no	southeast	1629.83

```
1336   21  female  25.8         0     no  southwest   2007.95
1337   61  female  29.1         0    yes  northwest  29141.36

[1338 rows x 7 columns]
```

```
scatter_matrix = pd.plotting.scatter_matrix(
    df_plot, figsize=(10, 10), alpha=0.5, marker='o', grid=True, s = 5
)

for ax in scatter_matrix.ravel():
    ax.set_xlabel(ax.get_xlabel(), fontsize=12)
    ax.set_ylabel(ax.get_ylabel(), fontsize=12)
    ax.xaxis.label.set_rotation(45)
    ax.yaxis.label.set_rotation(45)
    ax.yaxis.label.set_ha('right')

plt.suptitle("Scatter Matrix Plot", y=0.96, fontsize=16)

plt.show()
```



Insights

1. Expenses vs age - It is clear from the scatter plot that with age the medical expenses increases.
2. Expenses vs bmi - from the plot it can seen that there is a concentration of values and however there is not a clear trend but we can observe increase in expenses with increase in bmi.
3. Expenses vs children - There is no trend present between two features. Number of children has no major role in determining the expenses.

Observations

1. There are no missing datapoints in the given data.
2. We have already converted our categorical datasets into numerical datasets with the help of get_dummies.
3. Now moving towards train test split.

```
df_reg = data

for col in list(df_reg.columns):
  if str(df_reg[col].dtypes) == 'object':
    print(df_reg[col].unique())
```

```
['female' 'male']
['yes' 'no']
['southwest' 'southeast' 'northwest' 'northeast']
```

```
def cat_to_num(col_data, col_name, class_lis ):
  col_data[col_name] = col_data[col_name].apply(lambda x:
class_lis.index(x) + 1)

for cols in list(df_reg.columns):
  if str(df_reg[cols].dtypes) == 'object':
    cat_to_num(df_reg, cols, list(df_reg[cols].unique()))
```

For smoker: 1 = yes, 2 = no For sex: 1 = female, 2 = male For region: 1 = southwest, 2 = southeast, 3 = northwest, 4 = northeast.

```
X = df2.drop(['expenses'], axis = 1)
y = df2['expenses']

X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.3, random_state=42)
```

```
# Train the Linear Regression model
linear_reg = LinearRegression()
linear_reg.fit(X_train, y_train)

# Train the Ridge Regression model
ridge_reg = Ridge(alpha=0.5)
ridge_reg.fit(X_train, y_train)
```

```
# Train the Lasso Regression model
lasso_reg = Lasso(alpha=0.5)
lasso_reg.fit(X_train, y_train)
```

```
Lasso(alpha=0.5)
```

```
# Make predictions on the test set
linear_pred = linear_reg.predict(X_test)
ridge_pred = ridge_reg.predict(X_test)
lasso_pred = lasso_reg.predict(X_test)

# Calculate evaluation metrics
linear_mae = mean_absolute_error(y_test, linear_pred)
ridge_mae = mean_absolute_error(y_test, ridge_pred)
lasso_mae = mean_absolute_error(y_test, lasso_pred)

linear_mse = mean_squared_error(y_test, linear_pred)
ridge_mse = mean_squared_error(y_test, ridge_pred)
lasso_mse = mean_squared_error(y_test, lasso_pred)

linear_r2 = r2_score(y_test, linear_pred)
ridge_r2 = r2_score(y_test, ridge_pred)
lasso_r2 = r2_score(y_test, lasso_pred)

# Print the evaluation metrics
print("Linear Regression Metrics:")
print(f"MAE: {linear_mae}")
print(f"MSE: {linear_mse}")
print(f"R-squared: {linear_r2}")
print("\nRidge Regression Metrics:")
print(f"MAE: {ridge_mae}")
print(f"MSE: {ridge_mse}")
print(f"R-squared: {ridge_r2}")
print("\nLasso Regression Metrics:")
print(f"MAE: {lasso_mae}")
print(f"MSE: {lasso_mse}")
print(f"R-squared: {lasso_r2}")
```

```
Linear Regression Metrics:
MAE: 4144.88640999345
MSE: 33777093.10084606
R-squared: 0.7696351080608884

Ridge Regression Metrics:
MAE: 4148.229580129345
MSE: 33786028.61035601
```

```
R-squared: 0.7695741665323639

Lasso Regression Metrics:
MAE: 4145.170098628805
MSE: 33777925.44532053
R-squared: 0.7696294313454782
```

```python
coefficients_df = pd.DataFrame({
    'Feature': X.columns,
    'Linear Regression Coefficient': linear_reg.coef_,
    'Ridge Regression Coefficient': ridge_reg.coef_,
    'Lasso Regression Coefficient': lasso_reg.coef_
})

fig, ax = plt.subplots(figsize=(12, 4))
ax.axis('tight')
ax.axis('off')

table = ax.table(cellText=coefficients_df.values,
colLabels=coefficients_df.columns, loc='center', cellLoc='center')
table.auto_set_font_size(False)
table.set_fontsize(10)
table.scale(1, 1.5)
plt.show()
```

Feature	Linear Regression Coefficient	Ridge Regression Coefficient	Lasso Regression Coefficient
age	261.28251281367665	261.2340368153224	261.2818873443173
bmi	348.966009374454	348.90205476609844	348.8639840025136
children	424.4106794385628	424.61475040980724	424.1661768414394
sex__female	-52.49762358115285	-53.238056122754344	-103.34451397126217
sex__male	52.49762358116266	53.23805612286349	0.0
smoker__no	-11813.947297798173	-11794.7014672667	-23624.85561198286
smoker__yes	11813.947297798171	11794.70146726149	0.0
region__northeast	595.5377967043111	594.4653207189085	863.8101012568711
region__northwest	109.06784463070197	107.75297390655531	377.2136890076634
region__southeast	-375.08035908322427	-373.1307306124827	-102.56868556155607
region__southwest	-329.5252822517919	-329.0875640122228	-57.12941569160113

1. For α = 0.5 we are getting the lowest MAE ans MSE score and highest R^2 score for both ridge and lasso regression.
2. Intrestingly the lasso regression is making smoker_yes and sex_male 0, i.e., they are irrelevant features according to it but it is not true, both smoker_yes and sex_male show good correlation with expenses.

Observations

1. From the table it is clear that except sex_female, smoker_no, region_southeast, and region_southwest all have positive coefficents, which implies that these features will proportionately increase the expenses.

2. Based on sex, sex_male feature has positive coeffiecent whereas sex_female has negative coefficents implying that females have less medical expenses as compared to males.

3. Similarly, for people who are smokers have more medical expenses as compared to non-smokers, which can found in the nature of their coefficents, also the value of coefficent is large implying that it is a major feature.

4. Finally, region does not associate well with medical expenses as per the correlation but here we can see a person from northeast and northwest have more medical expenses than a person who is from either southeast or southwest.