Effect of Qualitative Predictors:

Let's say there are two columns – marital status and gender. The possible values in marital status are Married/Divorced/Single/Widowed and the possible values in gender variable are Male/Female/Other/Unknown.

How are the values in these variables going to affect the predictors: One approach is to use pivot tables. Other approaches include as shown below:

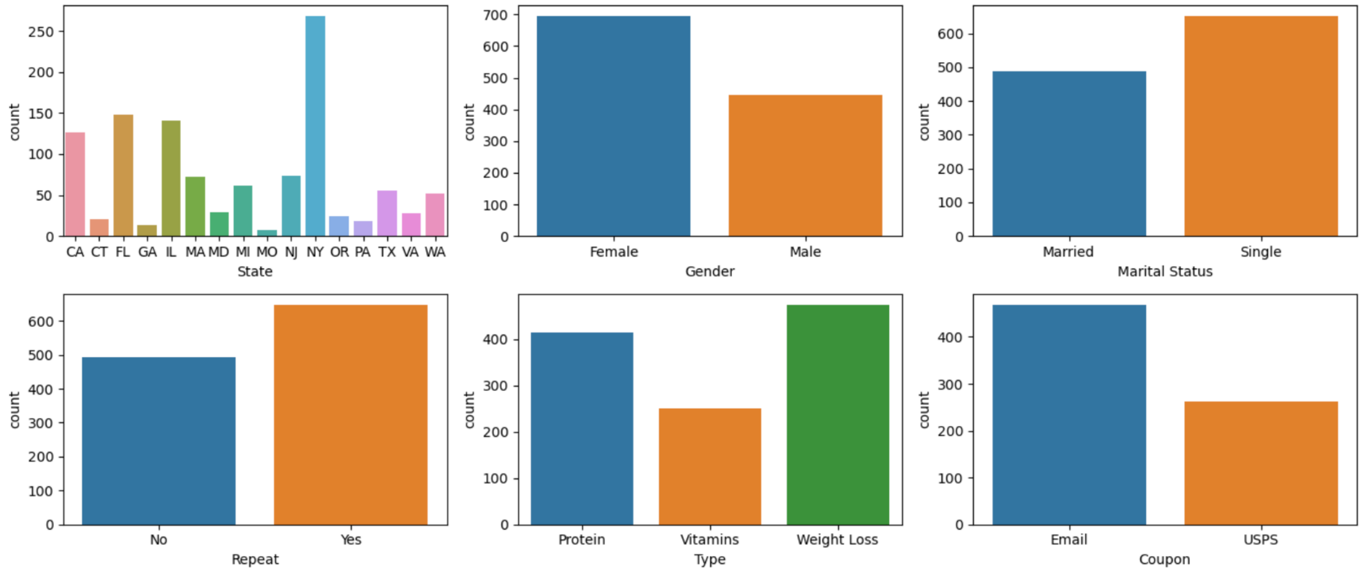
- Pivot tables become large and messy while working with multiple qualitative variables.
- Working with both qualitative and quantitative variables is not easy in pivot tables.
- Drawing statistical inferences is not possible using pivot tables.

Although pivot tables can be a useful tool for exploring datasets, usually, performing *linear regression using dummy variables should be preferred while modelling data*.

About Healthnutsonline dataset:

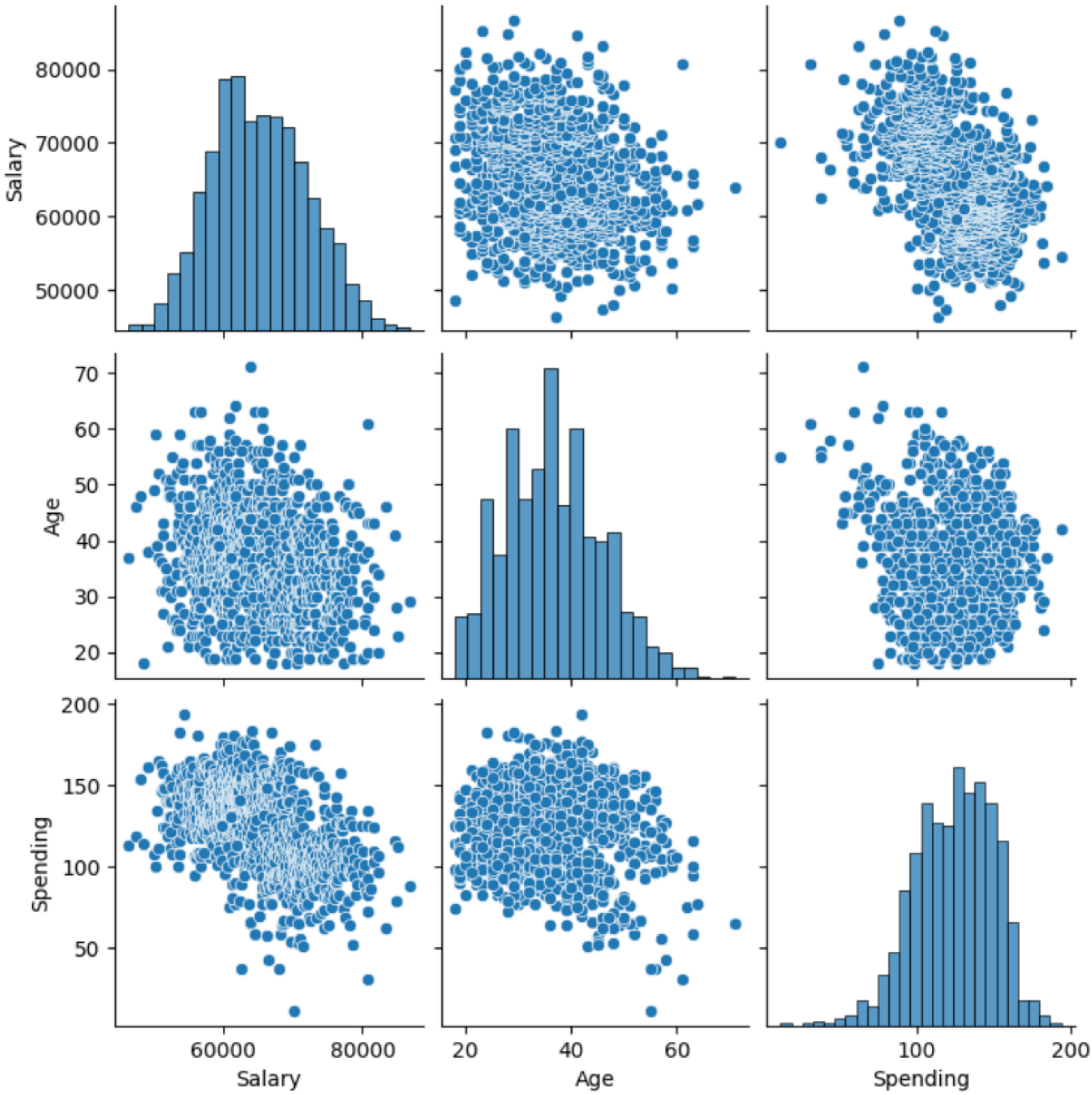
- Healthnutsonline sells vitamins, weight loss products and dietary supplements through its website.
- The dataset healthnutsonline.csv contains information on customers registered on their website.
- Customers also have an option of checking out as "Guest" without revealing any information about themselves. This data has been excluded.

Data Exploration:



- The distribution of orders for each state, gender, marital status is provided in the first row of the EDA diagram.
- The distribution of orders for customers who are either a repeat customer or not a repeat customer is given in the first diagram from the left in the second row.
- The distribution of the type of product is given in the second diagram from the left in the second row.
- The distribution if the coupon has been applied is given in the last diagram of the second row.

Scatter Plots



- As salary increases, spending seems to decrease.
- When age is increasing, salary is decreasing.
- The distribution for age is right skewed and for spending is left skewed. The salary distribution is normal.

Pivot Tables

Pivot tables are used to slice and dice the data. Below pivot table shows the avg spending for each gender.

```
table = pd.pivot_table(data = df, values = 'Spending', index = 'Gender', aggfunc = 'mean', margins = True)
table
```

Gender	Spending
Female	138.689597
Male	101.618812
All	124.186465

- Women spend on average $139.
- Men spend on average $102.

- Women spend $38 more than men on average.

How to do regression with qualitative predictors?

Single Category Predictors:

For example, to run regression of spending on gender, the gender variable needs to be recoded as a numeric variable. Typical to use a 0-1 coding as below:
1 – Male
0 – Female

The category 0 (Female) is called base or reference category. In Python, this is automatically handled when the variable is converted to a category variable using the astype() function.

In [37]:
```
# Create and train a linear regression model for the data and view its summary
# Note: The objective is to predict 'Spending' using 'Gender'
lr_model_1 = smf.ols("Spending~Gender",data=df)
lr_model_1 = lr_model_1.fit()
print(lr_model_1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               Spending   R-squared:                       0.498
Model:                            OLS   Adj. R-squared:                  0.498
Method:                 Least Squares   F-statistic:                     1130.
Date:                Mon, 01 Jan 2024   Prob (F-statistic):          1.28e-172
Time:                        15:17:23   Log-Likelihood:                -4922.4
No. Observations:                1140   AIC:                             9849.
Df Residuals:                    1138   BIC:                             9859.
Df Model:                           1
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept        138.6896      0.690    201.065      0.000     137.336     140.043
Gender[T.Male]   -37.0708      1.103    -33.616      0.000     -39.235     -34.907
==============================================================================
Omnibus:                       75.225   Durbin-Watson:                   1.993
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              121.575
Skew:                          -0.502   Prob(JB):                     3.98e-27
Kurtosis:                       4.246   Cond. No.                         2.44
==============================================================================

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

In the pivot table for spending for each gender, we observed that women spent $138 dollars and men spent $101 which is around $37 less than women. The coefficient in the above model is related to the pivot table. Below is the pivot table:

In [38]:
```
# Create a pivot table of mean 'Spending' with respect to 'Gender'
table = pd.pivot_table(data = df, values = 'Spending', index = 'Gender', aggfunc = 'mean', margins = True)
table
```

Out[38]:

	Spending
Gender	
Female	138.689597
Male	101.618812
All	124.186465

The estimated regression equation is:

Spending = 138.69 – 37.07 Gender

The same information is recoded in both the pivot table and model coefficients output.

- In general if X is a dummy variable (values 0 or 1) and the estimated regression equation is:

$$\hat{Y} = a + bX$$

- Then a is the average value of Y for the category X = 0 (i.e. the base or reference category).
- The average value of Y for the category X = 1 is $a + b$.
- Thus, b is the difference in average Y between the two categories.

Multi-Category Predictors

Example. Investigating amount spent on average for each coupon type. One way is to create a pivot table with the average spending for each coupon type. Below is the result of the pivot table:

```
In [27]: # Create a pivot table of mean 'Spending' with respect to 'Coupon'
         pd.pivot_table(data = df, values = 'Spending', index = 'Coupon', aggfunc = 'mean', margins = True)
```

Out[27]:

Coupon	Spending
None	124.526748
Email	128.926844
USPS	115.169618
All	124.186465

Another way is to create a regression model with Spending as the response variable and Coupon as the predicting variable. We could use dummy variables or let python set the values automatically.
We define the two dummy variables as below:

Coupon_E
1 = Email
0 = Otherwise

Coupon_U
1 = USPS
0 = Otherwise

In this system of coding,
- A customer who has received an email coupon has Coupon_E=1 and Coupon_U=0
- A customer who has received a USPS coupon has Coupon_E=0 and Coupon_U=1
- A customer who has received None has Coupon_E=0 and Coupon_U=0

The category for which all referen englishce variables are 0 is called base/reference category

In the above example, the reference category is None. In general, if there are m categories, we need m-1 variables. Any category can be made the base/reference category but that will change the interpretation of the coefficients.

```
In [26]: # Create and train a linear regression model for the data and view its summary
         # Note: The objective is to predict 'Spending' using 'Coupon' with 'None' as the reference category
         df['Coupon'] = df['Coupon'].cat.set_categories(["None","Email","USPS"])
         lr_model_5 = smf.ols("Spending~Coupon",data=df)
         lr_model_5 = lr_model_5.fit()
         print(lr_model_5.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               Spending   R-squared:                       0.043
Model:                            OLS   Adj. R-squared:                  0.041
Method:                 Least Squares   F-statistic:                     25.28
Date:                Tue, 02 Jan 2024   Prob (F-statistic):           1.81e-11
Time:                        10:32:23   Log-Likelihood:                -5290.7
No. Observations:                1140   AIC:                         1.059e+04
Df Residuals:                    1137   BIC:                         1.060e+04
Df Model:                           2
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
Intercept         124.5267      1.242    100.287      0.000     122.090     126.963
Coupon[T.Email]     4.4001      1.699      2.590      0.010       1.067       7.734
Coupon[T.USPS]     -9.3571      1.987     -4.709      0.000     -13.256      -5.458
==============================================================================
Omnibus:                       26.366   Durbin-Watson:                   0.990
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               27.711
Skew:                          -0.381   Prob(JB):                     9.61e-07
Kurtosis:                       3.054   Cond. No.                         3.71
==============================================================================
```

The estimated regression equation is:

$$Spending = 124.53 - 9.36 \times Coupon\_U + 4.4 \times Coupon\_E$$

In general,

$$Spending = a + b_1 \times Coupon\_U + b_2 \times Coupon\_E$$

- For a person with Coupon = None, the avg. spending is equal to a
- For a customer with coupon = USPS, the avg. spending is equal to $a + b_1$
- For a customer with coupon = Email, the avg. spending is equal to $a + b_2$

Multiple Regression

In [30]:
```python
# Create and train a linear regression model for the data and view its summary
# Note: The objective is to predict 'Spending' using 'Age' and 'Gender'
lr_model_8 = smf.ols("Spending~Age+Gender",data=df)
lr_model_8 = lr_model_8.fit()
print(lr_model_8.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               Spending   R-squared:                       0.580
Model:                            OLS   Adj. R-squared:                  0.579
Method:                 Least Squares   F-statistic:                     785.1
Date:                Tue, 02 Jan 2024   Prob (F-statistic):          6.57e-215
Time:                        12:12:36   Log-Likelihood:                -4821.0
No. Observations:                1140   AIC:                             9648.
Df Residuals:                    1137   BIC:                             9663.
Df Model:                           2
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept        170.3763      2.221     76.696      0.000     166.018     174.735
Gender[T.Male]   -41.1496      1.046    -39.342      0.000     -43.202     -39.097
Age               -0.8242      0.055    -14.878      0.000      -0.933      -0.715
==============================================================================
Omnibus:                        7.091   Durbin-Watson:                   2.039
Prob(Omnibus):                  0.029   Jarque-Bera (JB):                7.639
Skew:                          -0.132   Prob(JB):                       0.0219
Kurtosis:                       3.302   Cond. No.                         174.
==============================================================================

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

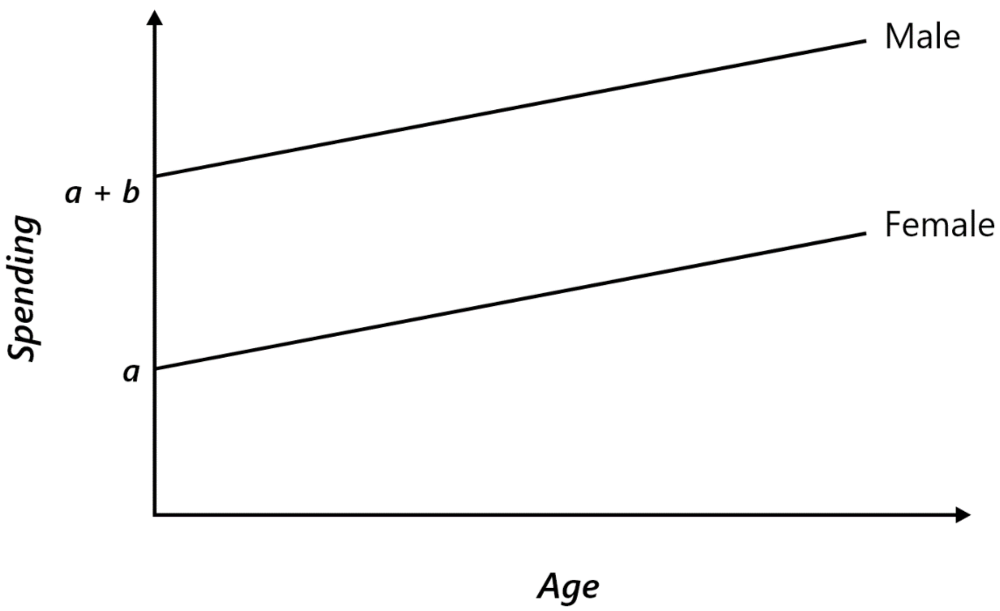
- Estimated regression equation in the output is:

$$Y = 170.38 - 41.15 \times Gender - 0.824 \times Age$$

- In general,

$$Spending = a + b_1 \times Gender + b_2 \times Age$$

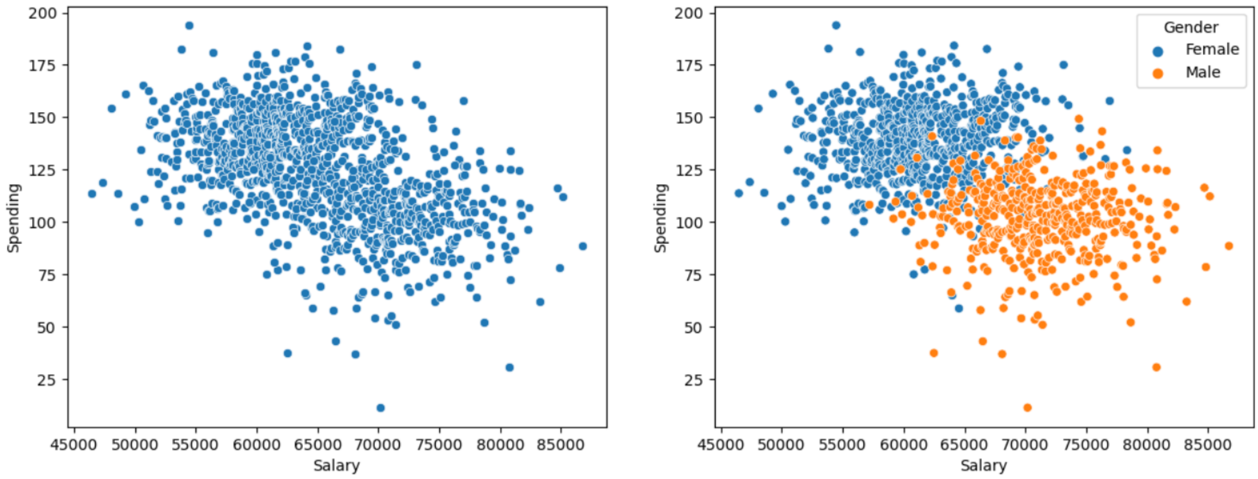
- For a female customer, $Spending = a + b_2 \times Age$
- For a male customer, $Spending = a + b_1 + b_2 \times Age$



In each of the cases, the intercept is different, but the slope is the same explained in above graph.

Therefore, a is of no economic significance. It is only EXTRAPOLATION!

In [30]:
```
# Create scatter plots of 'Spending' versus 'Salary', one colored by 'Gender' and the other without any categorical division
plt.figure(figsize = (14, 5))
plt.subplot(1, 2, 1)
sns.scatterplot(data = df, x = 'Salary', y = 'Spending')
plt.subplot(1, 2, 2)
sns.scatterplot(data = df, x = 'Salary', y = 'Spending', hue = 'Gender');
```



If gender is not accounted for in the above dataset, there seems to be a sharp decline in the salary. However, when the gender is accounted for, there is no such trend observed.