**Linear Regression using Qualitative Predictors**

**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.

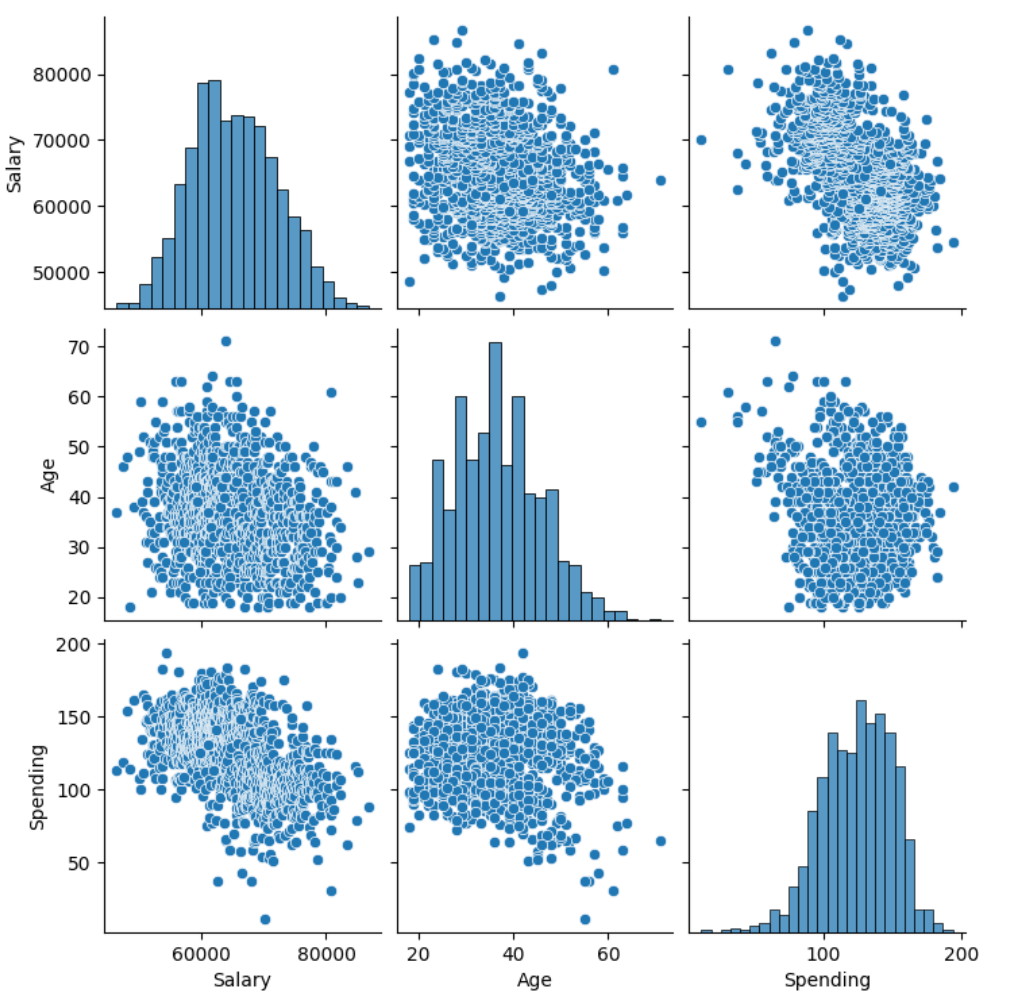
**Data Exploration:**

**A group of colorful bars

Description automatically generated with medium confidence**

* 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.

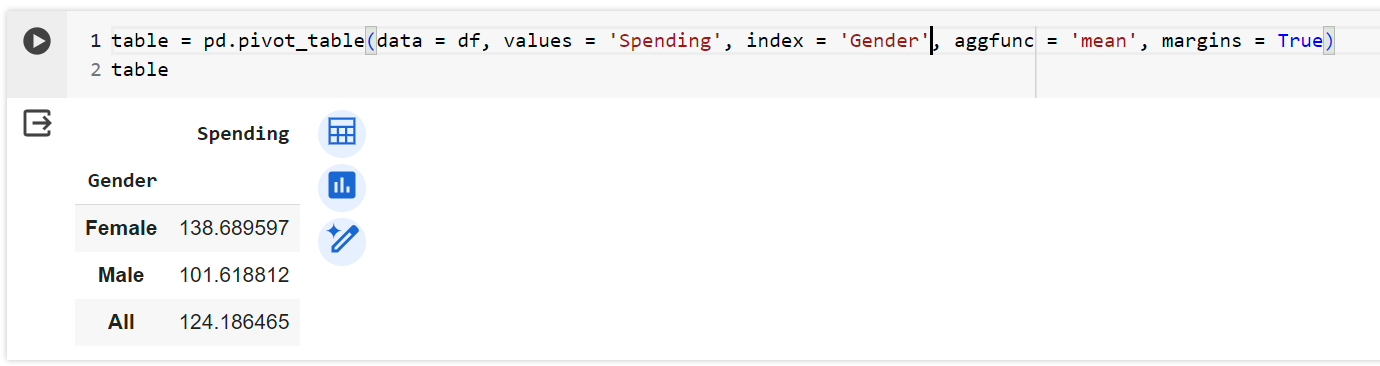
**Scatter Plots**



* 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.

****

* 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.

A screenshot of a computer

Description automatically generated

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:

A screen shot of a computer

Description automatically generated

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:
* 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 .
* 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:

A screen shot of a computer

Description automatically generated

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

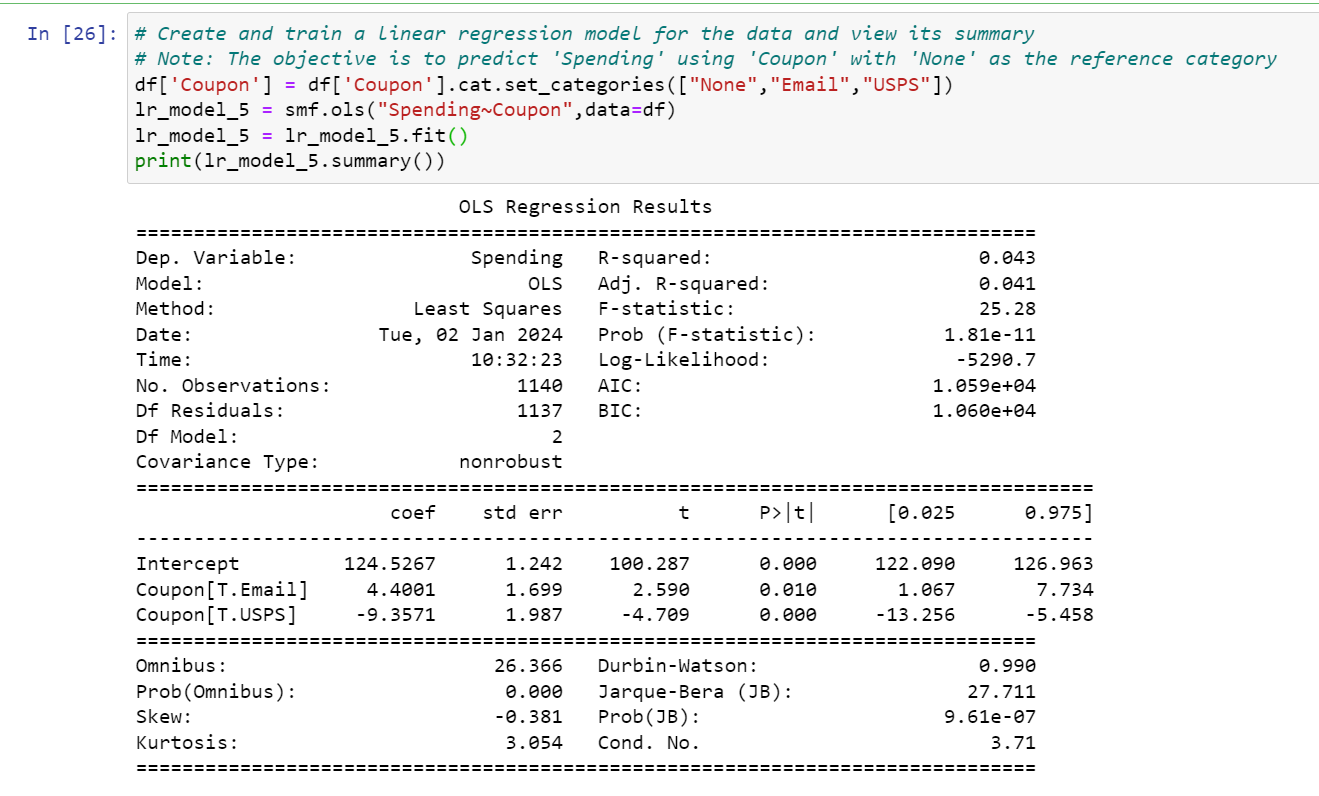
0 = Otherwise

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.

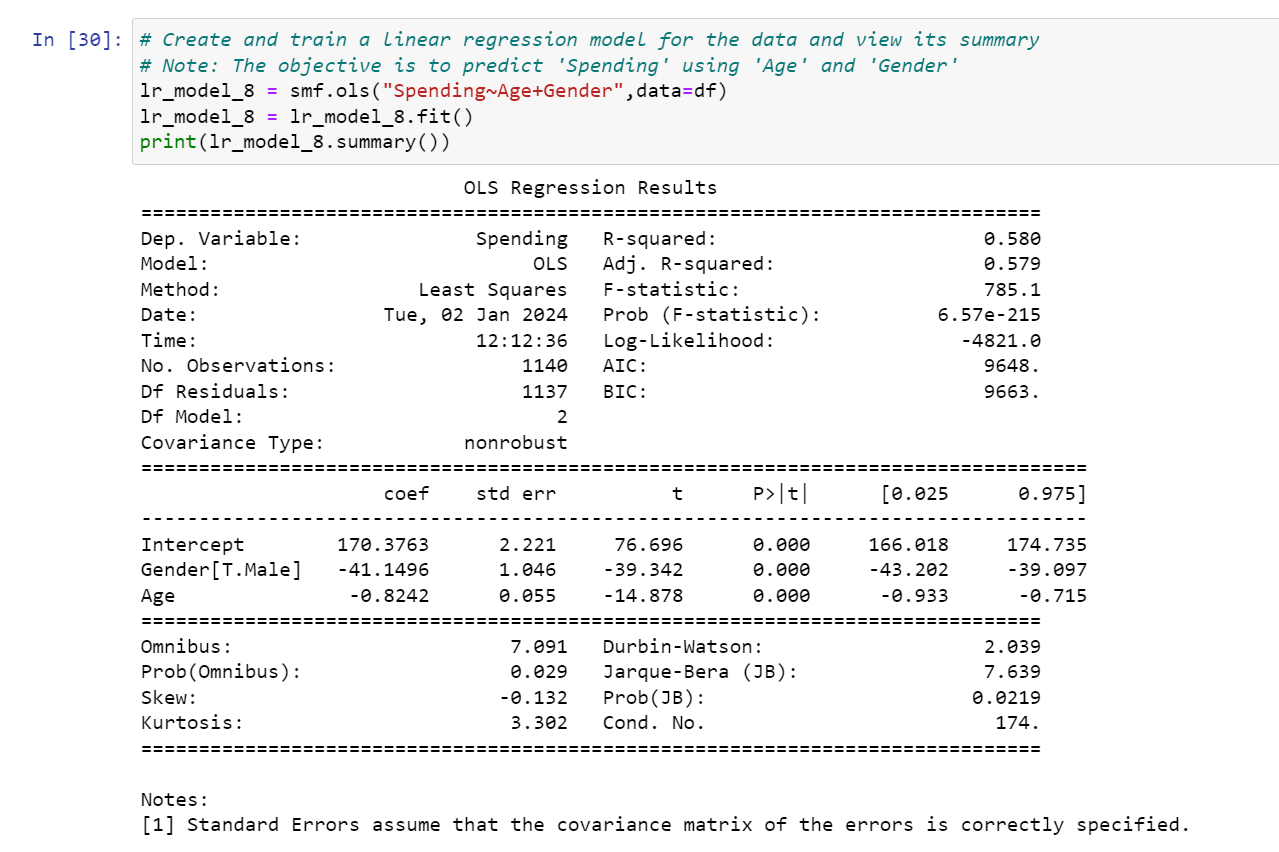


The estimated regression equation is:

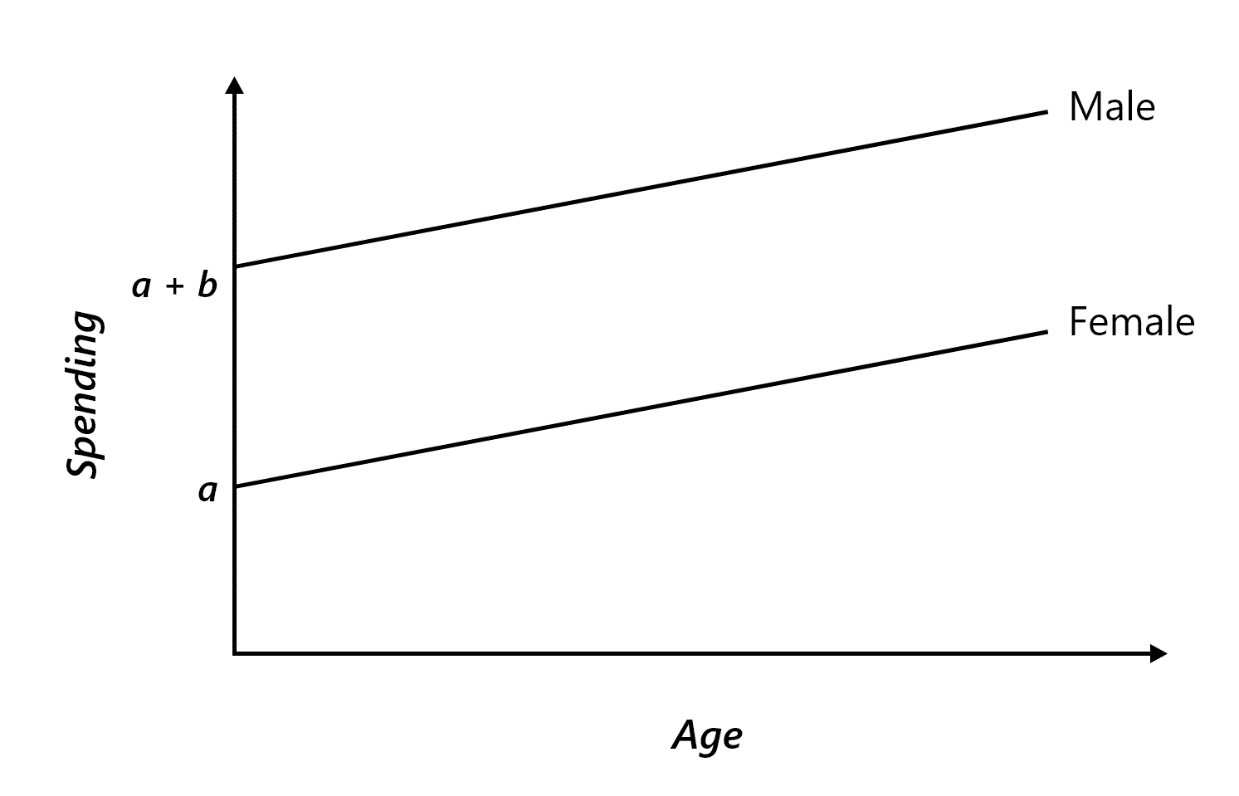
In general,

* For a person with Coupon = None, the avg. spending is equal to
* For a customer with coupon = USPS, the avg. spending is equal to
* For a customer with coupon = Email, the avg. spending is equal to

**Multiple Regression**

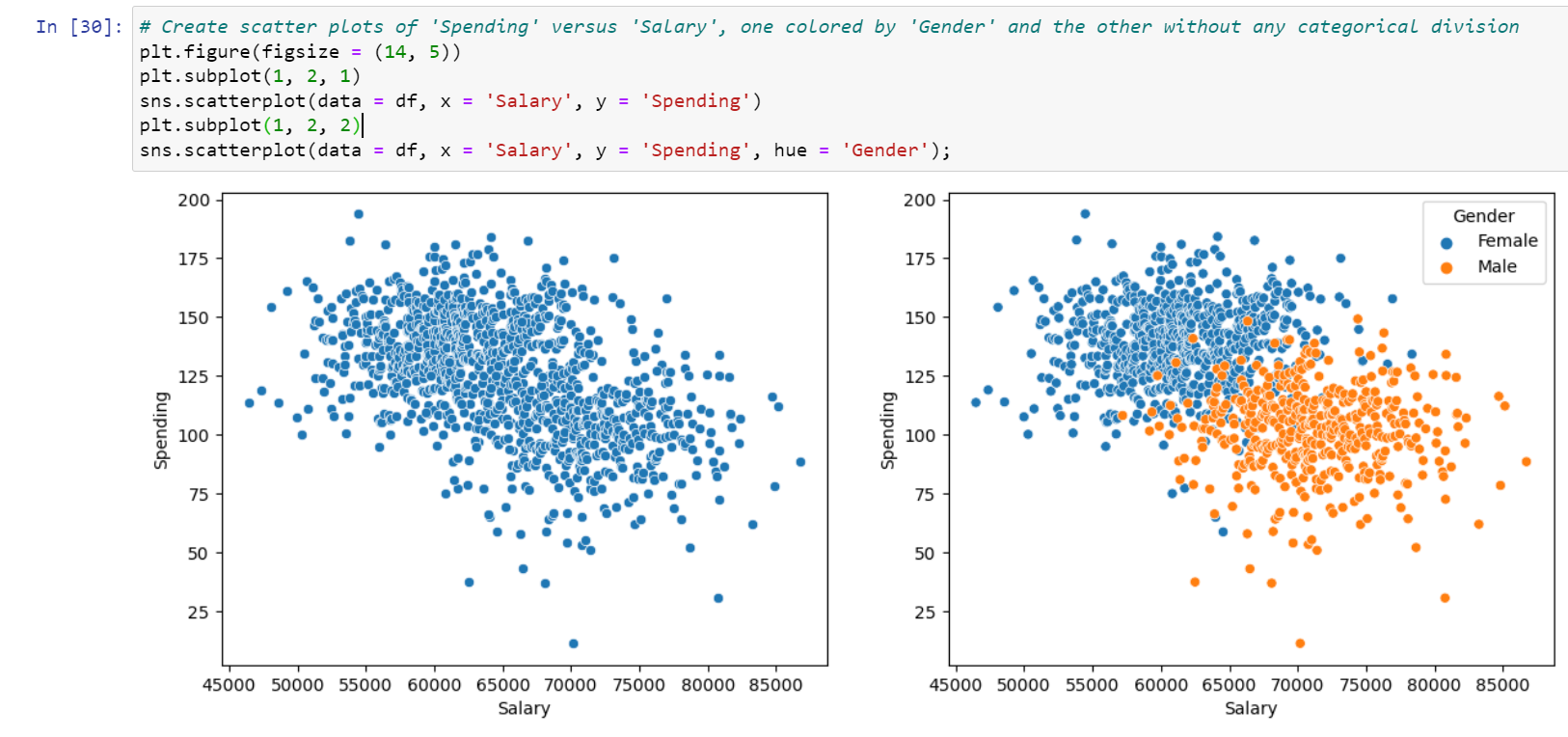
****

* Estimated regression equation in the output is:
* In general,
* For a female customer,
* For a male customer,



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!***



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.