Final Project Write-Up

To start, I chose to do a linear regression project because I enjoy predicting things based on certain outputs. I initially wanted to try to predict an NBA player's salary, but NBA data sets didn't have enough entries. That being said I used a data set from a class I took last year that has the data of the company's hardware department with all the info of each sales_rep. Like years in a company, female or male, NPS(rating), Feedback, College(IF), Personality, and salary. Based on these columns I wanted to predict the salary output. Salary is the intercept and the other columns are the coefficients.

First Module: Clean 1

In my first module, I knew I had to do some sort of data cleaning as the raw data had non-numerical values. Clean 1() does this perfectly. I take in the CSV and then create dummy variables for Personality so that three new columns are created with 1s and 0s. The college column also had yes or no which the function changes it to 1 and 0s.

Second Module: Clean 2

This module was done in the process as I started to create the linear regression code. I saw that I had extra columns and that salary wasn't the last or the first column indicating this would be a problem for my linear regression output. So I went back and made a second function called clean2. This function takes in the first modified CSV and drops unwanted columns like Sales_rep, personality, and college. It also makes the salary column the last one which I thought would make it easier to identify the Y of the equation.

Third Module: Graph (Runs fifth)

The third module was a graph. I got a lot of input form the TAs telling me that a graph could be a good addition to my project. Hence, with their help, I was able to create a graph. I chose NPS as the category I was going to analyze. NPS is the net Promoter Score for employees, which is a metric used to measure employee satisfaction and engagement within an organization. This means my expected output would be a line going up as NPS increases salary also increases. At first, I had problems with this, but what I did was that I created a function that averages the salary at each NPS meaning the line would be linear and it worked perfectly. My outputs was a line upward sloping.

Fourth Module: LR

Here I used my knowledge of a QM modeling class and lectures from 210. I knew that to create a linear equation to predict salary I needed to convert each factor of salary into coefficients to implement in the equation. From research I learned that I had to use linfa_linear to do this. In my function I introduced the y = salary and than the x = all the columns besides salary.

This indeed gave me the intercept and parameters to use for the linear regression equation in the other module. This was probably the most difficult module as I had to do a bunch of things like wrap and convert features and targets. The output was good and made sense as I compared it the output an excel gave me.

Fifth Module: LE

For the fifth module, this was simple. I had used the output from M4 and created the linear equation using the function Le. I also added prompts so that anyone can implement their credentials and see their expected salary.

Four prompts are binary:

I tried my best to specify them in the output.

For all 1= yes and 0= no

1. College
2. Explorer
3. Diplomat
4. Analyst

Some things have to be taken into account to make the equation reliable:

1. You can only be one of the three personalities.
2. NPS Max = 10
3. Feedback Max = 4

Main Module:

I made every function into Mods so that anyone can run it. You just run the main.

Your Outputs:

1. Cleaned CSV 1
2. Cleaned CSV 2
3. Intercept and Coefficients
4. Prompts and Predicted Salary
5. Graph
6. Test 1
7. Test 2

Test 1:

Test 1 verifies that if every value is = 0 the output is the intercept meaning that the equation follows the properties interned of intercept = age(0) + x(0) + x (0)

Test 2:

For test two I had the idea of testing out if my equation was reliable in the sense that higher inputs would end up with higher salaries or in other words are more qualified person would output a higher salary which indeed it did.