Juan Vargas

Assignment 1

1. Code is in https://github.com/juanvargas7/ML/blob/main/blood_preassure.ipynb

a.

$$\text{systolic\_bp\_after\_treatment} = -7.4072 + 1.0224 \cdot \text{initial\_bp} - 3.0678 \cdot \text{drug\_dosage} + 0.0583 \cdot \text{age} + 2.7864 \cdot \text{sex}$$

$$systolic\ blood\ preassure\ after\ treatment = -7.4072 + 1.0224 * initial\ blood\ preassure - 3.0678 * drug\ dose + 0.0583 * age + 2.7864 * sex + \epsilon$$

b. Based on the statistical metrics; our model appears to be effectively predicting blood pressure post-treatment. For instance, it boasts a relatively low RMSE of 3.28. Moreover, the adjusted R-squared value suggests that with the selected variables and covariates, we can account for 96% of the variance. Examining the diagnostic plots (refer to Fig. 1, 2, and 3) offers further insights: in Fig. 1, the residuals are not randomly scattered, and the Q-Q plot in Fig. 3 reveals non-normality in the tails, with slight deviations around the center. The patterns observed, especially the parabolic shape in the residuals plot, suggest that a second-order polynomial regression might provide a more accurate fit.

```
RMSE: 3.2819894286236315
                            OLS Regression Results
==============================================================================
Dep. Variable:     systolic_bp_after_treatment   R-squared:                       0.960
Model:                                     OLS   Adj. R-squared:                  0.960
Method:                          Least Squares   F-statistic:                     1489.
Date:                         Tue, 19 Sep 2023   Prob (F-statistic):          1.48e-170
Time:                                 16:43:47   Log-Likelihood:                -651.85
No. Observations:                          250   AIC:                             1314.
Df Residuals:                              245   BIC:                             1331.
Df Model:                                    4
Covariance Type:                     nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          -7.4072      3.216     -2.303      0.022     -13.741      -1.073
initial_bp      1.0224      0.015     70.299      0.000       0.994       1.051
drug_dosage    -3.0678      0.133    -23.091      0.000      -3.329      -2.806
age             0.0583      0.024      2.386      0.018       0.010       0.107
sex             2.7864      0.422      6.610      0.000       1.956       3.617
==============================================================================
Omnibus:                       47.063   Durbin-Watson:                   1.901
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               73.884
Skew:                           1.076   Prob(JB):                     9.04e-17
Kurtosis:                       4.570   Cond. No.                     2.63e+03
```

Juan Vargas

Figure 1

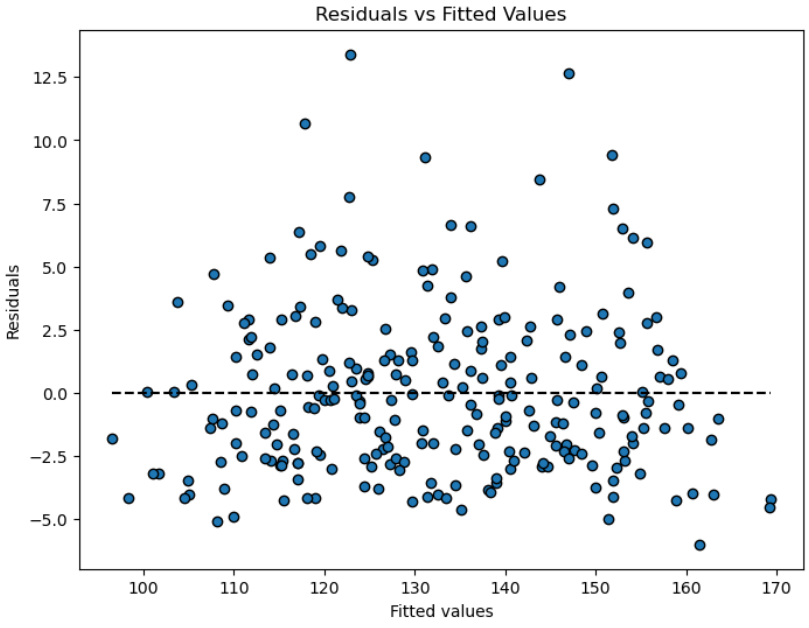


Figure 2

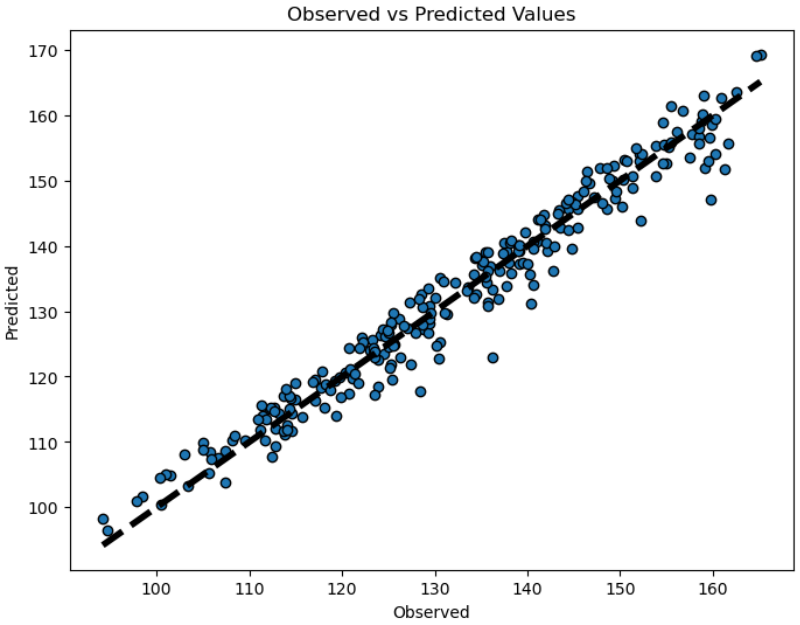
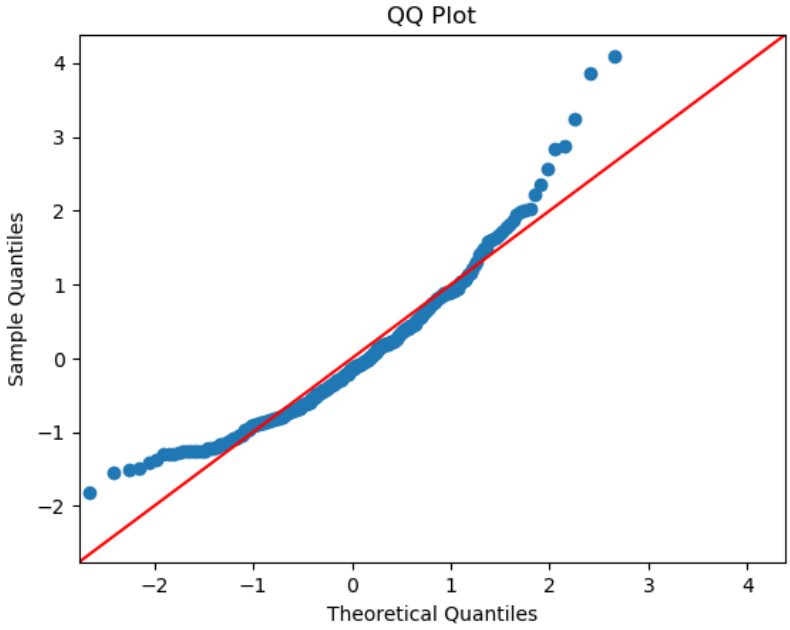
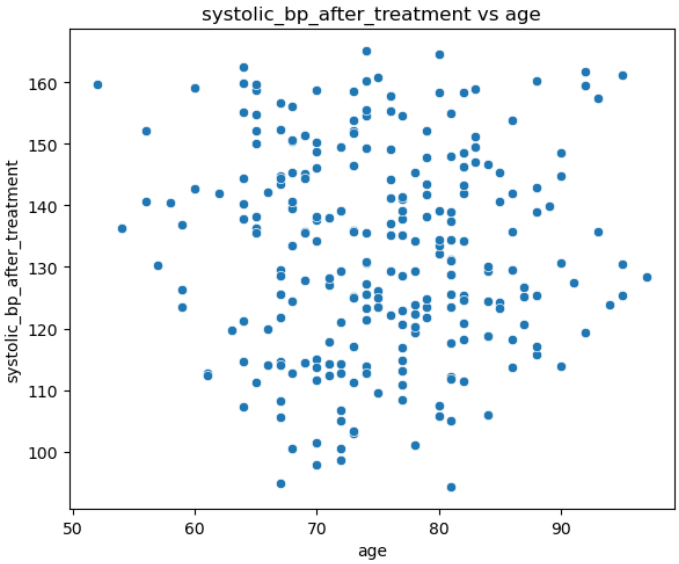


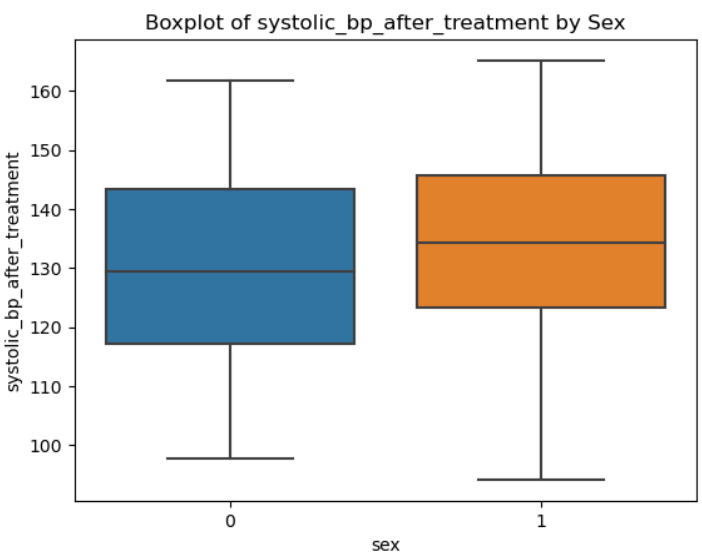
Figure 3

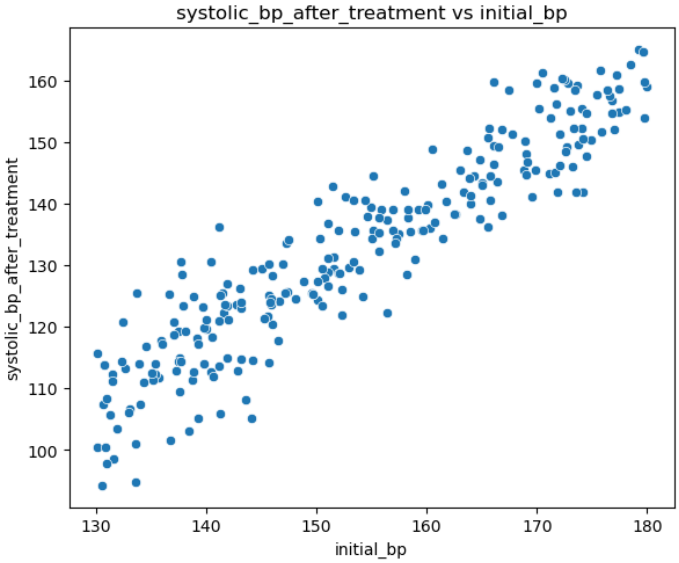


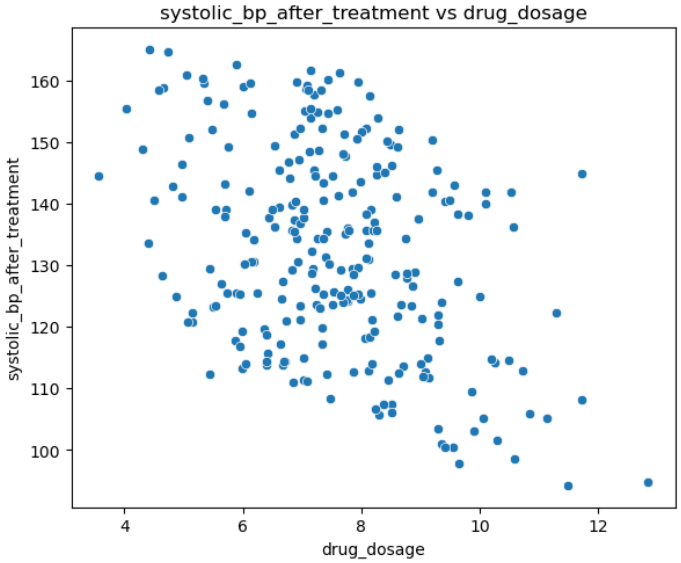
Juan Vargas

2.









After evaluating the correlation of -0.010945 and examining the plots, it's evident that age doesn't demonstrate a robust linear correlation with the outcome. Although, within the model, the p-value for age stands at 0.018. This suggests that the coefficient for age enhances the model beyond the mere inclusion of the intercept, meaning a significant linear relationship between age and the outcome. In contrast, the boxplot for sex reveals no discernible difference between genders. Yet, the model indicates a p-value of 0 for sex, emphasizing that its coefficient offers improvement over solely using the intercept. The initial blood pressure showcases a pronounced linear correlation, evidenced by a correlation value of 0.931311. Correspondingly, the model reports a p-value of 0, further solidifying its linear relationship with the outcome. Finally, drug usage appears to maintain a negative linear relationship, as suggested by a correlation of -0.391196. The model reinforces this observation with a p-value of 0, underscoring a notable linear relationship between drug usage and the outcome.