**HMW1**

**a.** From the scatterplot and correlation matrix, we see the following: Satisfaction among patients and Age have a strong negative correlation, i.e., Satisfaction among patients decreases fairly linearly as Age of the patients increases. Satisfaction among patients and Severity have a negative correlation, i.e., Satisfaction among patients decreases somewhat linearly as Severity of the condition in patients increases. Also, satisfaction among patients and Anxiety have a negative correlation, i.e., Satisfaction among patients decreases somewhat linearly as Anxiety among the patients increases

**b.** AGE-SEVERITY: They show a weak positive correlation with Age somewhat increasing linearly with Severity.

AGE-ANXIETY: They show a weak positive correlation with Age somewhat increasing linearly with Anxiety.

SEVERITY-ANXIETY: They show a strong positive correlation with Severity increasing fairly linearly with Anxiety.

**e.** Without Intercept: When we fit the linear regression model without an intercept, we see that the coefficients of all the three predictor variables are negative which follows the correlation matrix data, but the R² value of the model comes out to be negative, which means that the fit explained by this model is actually worse than the fit explained by the horizontal mean value line for the predictor variables. This practically means that this model probably has an intercept other than (0,0) through which the line passes and for which the sum of squared error is minimum.

With Intercept: When we fit the linear regression model with an intercept, we see that the coefficients of all three predictor variables are negative which follows the correlation matrix data, also, the R² value for this model is 0.673 which means that the regression line explains 67.3% of the variance of predictor variables and is able to fairly explain the datapoints in the scatterplot diagram.

**f.** Residuals vs Fitted Values: As we can see from the plot, the points for the residuals vs the fitted values are fairly random and scattered without displaying any serious relationship between the two values.

Residuals vs Predictor Variables: As we can see from the plot, the points for the residuals vs the predictor variables are fairly random and scattered without displaying any serious relationship between the two values.

Residuals vs Each Two Factor Interaction: As we can see from the plot, the points for the residuals vs each two factor interaction of response-predictor as well as predictor-predictor variable are fairly random and scattered without displaying any serious relationship between the two values.

Normal Probability Plot for Residual: We can see from the normal probability plot for Residuals that the data points are not distributed around the standard normal distribution line with many outliers. This implies that the Residuals do not follow normal distribution and the points are located away from the mean.

Normal Probability Plot for Patient Satisfaction: We can see from the normal probability plot for Patient Satisfaction that the data points are fairly distributed around the standard normal distribution line with a few outliers. This implies that the data for Patient Satisfaction can be said to be a fairly normal distribution.

Normal Probability Plot for Patient Age: We can see from the normal probability plot for Patient Age that the data points are more distributed in an inverted C pattern around the standard normal distribution line with some outliers. This implies that the data for Patient Age does not follow Normal distribution and is a right-skewed distribution.

Normal Probability Plot for Patient Anxiety: We can see from the normal probability plot for Patient Anxiety that the data points are not distributed symmetrically around the standard normal distribution line with a fairly higher number of outliers. This implies that the data for Patient Anxiety does not follow Normal distribution and is a right-skewed distribution as there are more outliers under the standard normal distribution line.

Normal Probability Plot for Severity: We can see from the normal probability plot for Severity that the data points are fairly distributed uniformly around the standard normal distribution line with just one outlier. This implies that the data for Severity fairly follows Normal distribution.

**g.** After running ANOVA test on the dataset, we find the estimated coefficients, standard errors associated with these coefficients, t-statistic value, F-statistic value and the p-value of the predictor variables.

Coefficients: All predictor variables have negative coefficients meaning the predictor variables are negatively correlated with Patient Satisfaction and as the values of predictor variables increases, the value of Patient Satisfaction decreases.

Standard Error: The SE for Age and Severity are low while it is higher for Anxiety which shows that under repeated sampling, Age and Severity vary slightly while Anxiety of a patient varies significantly under repeated sampling.

t-Statistic and P-Value: The t-statistic values for Age, Anxiety and Severity are negative. This means that most of the data points of Age, Anxiety and Severity lie to the left of the mean of the whole data set. From the p-values, we observe that only Age has a considerably lower value from α (0.05) and thus we can reject the null hypothesis and conclude that there is a statistically significant difference between these values and Patient Satisfaction values while for Severity and Anxiety we cannot say the same due to its higher p-value than α, thus, we cannot reject the null hypothesis and conclude that there is no statistical difference between the values of Anxiety and Patient Satisfaction.

Omnibus: The higher value of Omnibus from zero indicates the dataset deviates from a normal distribution.

Prob(Omnibus): This value is also not very close to 1 which further strengthens our assumption that the dataset is not normally distributed.

Skew: The negative value of skew indicates that the distribution of residuals is negatively skewed which means the left hand tail is longer than the right hand tail and residual mean lies towards the right of the ideal mean.

Kurtosis: The positive value of kurtosis indicates that the residuals are not tightly clustered around zero and there are significant outliers.

Condition Number: The high of value of this parameter denotes that the output (Patient Satisfaction) is highly sensitive of the input variables (Age, Severity, Anxiety).

**h.** The p-values obtained from ANOVA indicate that there is a relationship between Patient Age and Satisfaction, while for Anxiety and Severity we cannot say the same due to their higher p-values.

**i.** As we can see from the correlation matrix, the correlation factor for Age is the highest with Patient Satisfaction followed by Anxiety and Severity. Hence, we can say that Age has a strong relationship with Patient Satisfaction followed by Anxiety and then lowest for Severity.

**j.** From the ANOVA results, we can say that all the three predictor variables have a negative relationship with Patient Satisfaction.

**k.** Confidence Interval:

β0 = From the calculations for confidence interval for β0 of individual predictor variables, we observe that Severity has the biggest confidence interval followed by Anxiety and Age. This means that the interval in which there is a 95% chance for the true value of β0 to lie in is the largest for Severity then Anxiety and at last is Age.

β1 = From the calculations for confidence interval for β1 of individual predictor variables, we observe that Anxiety has the biggest confidence interval followed by Severity and Age. This means that the interval in which there is a 95% chance for the true value of β1 to lie in is the largest for Anxiety then Severity and at last is Age.

**l.** X² Transformation: By applying this transformation on various variables, we observe that the data points become larger and more distributed and scatter further away from the mean and thereby increasing the standard deviation of the dataset.

X^½ Transformation: By applying this transformation on various variables, we observe that the data points become smaller and are more closely distributed around the mean and thereby decreasing the standard deviation of the dataset.

Log(X^½) Transformation: By applying this transformation on various variables, we observe that the data points become further smaller and are more closely distributed around the mean and thereby decrease the standard deviation of the dataset even more.

As we apply these transformations in this sequence, we notice that the data points align themselves more and more in a linear fashion.

**m.** Upon adding polynomial features and interactions up to degree 3, we notice that the xᶾ coefficients for Age, Anxiety and Severity are significant, especially, Anxiety. The xᶾ coefficient Anxiety and Severity has a negative sign, while Age has a coefficient with a positive sign. Hence, all three interactions between Patient Satisfaction and Age, Anxiety and Severity are significant. This statement is further strengthened by the R² value which is significant (0.992). Hence, a model of a polynomial with degree 3 better explains the relationship between predictor and response variable from a low variance point of view, but if we trade variance for bias, the linear regression was a better suited model for the response-predictor interaction. As we can see from the ANOVA, Age and Anxiety show a small p-value, hence there is an evidence of non-linear relationship between Patient Satisfaction and Age and Anxiety, while for Severity, the p-value is quite high (>α) and so we can say that there is no non-linear relationship between Satisfaction and Severity.