RESULTS

Task 1. Basic Statistics Analysis

1. We plot the histogram for each feature and we see that the distribution for each variable differs. Visually, we can say that X1 and X4 follow normal distribution whereas Y follows a gamma distribution. We do not see any variable following a particular type of probability distribution strictly. We plot QQ plots to examine this further.
2. To remove outliers, I use z-score. The **Z-score** is the signed number of standard deviations by which the value of an observation or data point is above the mean value of what is being observed or measured. I use a threshold of **3** to remove outliers.
3. I plot the correlation matrix in 2 ways for better analysis of the matrix. The correlation matrix indicates that there is a very strong correlation between X1 and Y. A feature highly correlated with output is a good candidate for a predictor. The next correlation that we see is between X4 and Y, but this correlation is not as strong as the previous one.

Task 2: Simple Linear Regression

1. For simple linear regression, OLS is used. *Ordinary least squares (OLS) regression* is a statistical method of analysis that estimates the relationship between one or more independent variables and a dependent variable; the method estimates the relationship by minimizing the sum of the squares in the difference between the observed and predicted values of the dependent variable configured as a straight line.
2. $a_0 = -3770.1268,\ a_1 = 256.5985$, variance = 1603867.24
3. A low **p-value** (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to the model because changes in the predictor's value are related to changes in the response variable. We have a p-value of 0 which indicates that it is statistically significant.
4. **Regression coefficients** represent the mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant. We have a coefficient of 256.5985 which indicates that for every additional increase of 1 unit in X1, we expect Y to increase by 256.59
5. **R-squared** is a statistical measure of how close the data are to the fitted regression line. tHigher the R-squared, the better the model fits the data. We get a R-squared value of 96.3% which indicates that the model fits the data very well.
6. **The "F value" and "Prob(F)"** statistics test the overall significance of the regression model. Specifically, they test the null hypothesis that *all* of the regression coefficients are equal to zero. The F value is the ratio of the mean regression sum of squares divided by the mean error sum of squares. With a low value of Prob (F-statistic):

1.12e-205, we can reject the null hypothesis and conclude that the regression equation does have some validity in fitting the data.

7.

a. Residual analysis

Our residual QQ plot is not a straight line. A qq plot plots the quantiles against each other. If we get a straight, diagonal line in the positive direction (i.e. a slope of 1), this indicates that the quantiles match the quartiles of a "theoretical normal distribution." Since the line is curved, our residuals do not follow a normal distribution. There is also literature that indicates that real-world normally distributed data does not show a typical normal distribution and can deviate a little and can be skewed. But from the QQplot and the histogram, we conclude that it is not normal.

b. Residual histogram

A residual histogram is plotted such that each bin has a minimum of 5 values. The last bin is not of equal width with the others. The histogram also indicates that the residuals do not follow a normal distribution

c. Chi Square test

I performed a chi-square test in 2 methods. One uses the CDF as mentioned in the textbook whereas the other uses a scaling factor and the middle value of the bins to calculate the expected frequency. We see that both the results of the chi-square statistic give different values. Based on reading and observation, we see that the chi-square test is not a very powerful test to check normality and is not dependable. The results vary based on bin size, bin width, value counts per bin and also depend on the method used to calculate expected frequencies.

The chi-distribution table gives a value of 0.71 for degree of freedom =2 and alpha=0.95. The null hypothesis is that the residuals follow the normal distribution.

The null hypothesis is accepted if $\chi^2 < \chi_{0.95,l-c}$.

$\chi^2 obtained = 55.9$ *and* $\chi_{0.95,l-c} = 0.71$

This implies that the residuals don't follow a normal distribution.

8. Scatterplot

When a pattern is observed in a residual plot, a linear regression model is probably not appropriate for the data, suggesting a better fit for a non-linear model.

9. Using a higher degree polynomial, does not improve our model. We do not see any significant improvement from our previous model. The p-value indicates that "X2" is significant, but the f-statistic and R-squared value indicates that it does not significantly improve our model.

Task 3. Linear Multivariable Regression

1. $a_0 = -1.147e+04,\ a_1 = 256.2235,\ a_2 = 5.0065,\ a_3 = 7.3778,\ a_4 = 7.1358, a_5 = 6.6433$,
   Variance : 1477124.01
2. We remove "X2" as the p-value>0.05. We look at r-squared and F- value and see that they show significant results after removing "X2". Even though X3 shows very less correlation with Y, it does increase R-squared by 0.001% and I retain it in the model.
3. The QQ-plot indicates that the residuals do not follow a normal distribution.
4. **Residual histogram**

   A residual histogram is plotted such that each bin has a minimum of 5 values. The histogram also indicates that the residuals do not follow a normal distribution
5. **Chi Square test**

   I have once again used 2 methods of getting a chi-square statistic and we see varying values. The chi-distribution table gives a value of 0.71 for degree of freedom =2 and alpha=0.95. The null hypothesis is that the residuals follow the normal distribution.

   The null hypothesis is accepted if $\chi^2 < \chi_{0.95,l-c}$.

   $\chi^2 obtained = 153.9\ and\ \chi_{0.95,l-c} = 0.71$

   This implies that the residuals don't follow a normal distribution.
6. **Scatterplot**

When a pattern is observed in a residual plot, a linear regression model is probably not appropriate for the data, suggesting a better fit for a non-linear model.

7. **Omnibus/Prob(Omnibus)** – a test of skewness. We hope to see a value close to zero which would indicate normalcy. The Prob (Omnibus) performs a statistical test indicating the probability that the residuals are normally distributed. We hope to see something close to 1 here. In this case Omnibus is high indicating that the residuals are not normal.
8. **Skew** – a measure of data symmetry. We want to see something close to zero, indicating the residual distribution is normal. This result has a high skew and hence not normal.