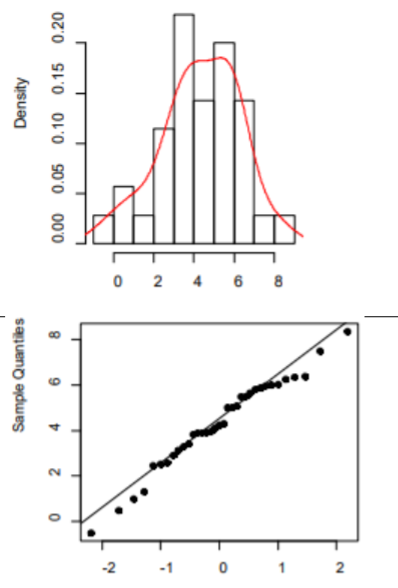
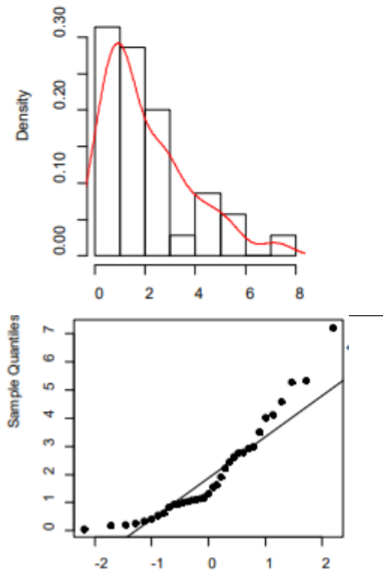
Question 1:

1. As given in the question, a normal distribution peaks in the middle and is symmetric about the mean which means it is a symmetric distribution that is a distribution which the left and right sides of the distribution are mirror images of each other. A symmetric distribution by definition is never skewed.

Explaining in the terms of skewness of the data: it is a mathematical or statistical computational method for measuring a distributions or dataset’s unevenness. It describes the distribution of the plurality of data values about the mean. From the given question we can say that if it a data is normally distributed and symmetrical then the coefficient of skewness is equal to zero or approximately near zero that is zero skewness.

Explaining in terms of kurtosis of the data: it is a statistical approach that tests the sharpness of the peak in the distribution of data. From the given question, if the data is normally distributed and is symmetrical then the kurtosis coefficient is qual to or equivalent to 3, and the distribution of data us called mesokurtic.

1. Given, two datasets tests against normality under density comparisons and QQ-plots and we have to describe whether the data appeared to be normal or approximately normal, skewed or any other distribution.

Figure(a) figure(b)

According to question and figure(a),we notice that the distributed data appears to be a straight line roughly, so we conclude that the distribution of the data is drawn from a normal distribution, it is a symmetric distribution, that means we consider that it has no skew or very less skewness, that is we can say that the mean is approximately equal to the median.

From the figure(b), we can conclude that the data appear to be a right skewed distribution. It is a distribution which is non-symmetric and have a long tail that points to extreme values on the distribution’s right edge which means the mean is higher than the median. Its is a positive skew data set of distribution.

1. A hypothesis testing is a inference about the data obtained for some experiment, i.e. it is mainly based on the assumptions made. The presumptions does not have to be correct all the time. In certain ways, hypothesis checking is a systematic method of validating the hypothesis. Hypothesis checking is used to determine the legitimacy of a proposition or observation made regarding the larger population. Null hypothesis states normal distribution whereas alternative hypothesis states the opposite of null, if null hypothesis is rejected, we assume it as alternative hypothesis that is not normally distributed.

Stating the hypothesis in terms of p-value:

1. A p-value is ultimately used in hypothesis testing to weigh the reliability of the proof, it can be shown in the following way- a p-value less than 0.05 suggest clear proof against null hypothesis, therefore it is rejected.
2. A p-value higher than 0.05 means there isn’t insufficient evidence to reject null hypothesis, so we accept null hypothesis.

Null hypothesis: the data distribution is normal i.e. p-value>0.05, we assume normality. A Shapiro-Wilk test is performed, where we get the p-value for both the data set. The distorted data has a p-value of 0.0016, indicating clear proof of non-normality. Since p-value is 0.o5847 for the roughly normal distributed results, the null hypothesis is held at the 05 percent level of significance. As a result, normally can be assumed for the data\_set\_1 and an effective parametric test cab be used if all other test assumptions are fulfilled.

1. When the data set or the distribution are not normally distributed, a data scientist will have to perform a Non-parametric test, it is a test where they do not approximate or assume model parameters using a normal distribution. We can also change the dependable variable by taking log or square root of the dependable variable. There are some non-parametric tests which can be performed: Mann-Whitney, Wilcoxon Signed Rank Test, Kruskal Wallis, Spearman’s test and Friedman Test.

While non-parametric tests entail fewer assumptions and can be applied to a broader variety of data types, parametric tests are favoured because they are more susceptible to identifying variations between samples or an influence of the independent variable on the dependable variable which means that when the data is usually distributed, a greater sample size is needed for the non-parametric test to detect any effect at a defined significance level than for the analogous parametric test.