Linear modelling using R

Steps-

1. Set working directory.
2. Copy source file into the working directory.
3. Import the source file into R. (load it into a dataframe)
4. Use attach () command for easy access of the loaded dataset.
5. Check the imported file if it has been imported correctly.
   1. File name – Sales\_data.csv (contains 200 records)
   2. Get the metadata about imported file using str(dataframe) function.
6. Validate and clean the data.
   1. There are a total of 200 observations with 4 variables.
   2. Attractiveness is a continuous numeric variable, while sales, plays and advertise budget are discrete integers.
   3. Attractiveness seems like a factor as it has reparative digits ranging from 1 to 10.
   4. The data looks clean, no invalid values.
7. Do descriptive analysis using summary() command-
   1. Advertise budget-
      1. The minimum advertisement budget is 9 while 2272 is the most.
      2. The average budget is 608 per advertise. A density plot shows that 2272 is an outlier which is the highest budget.
      3. It is not a normal distribution; it shows a positive skewness with tail at right side.
      4. It shows a positive Kurtosis, the probability of extreme values is lower than the normal distribution.
   2. Sales-
      1. Sales are ranging from 10 to 360 with average sales of 193.2.
      2. A density plot shows that Sales have bi-nodal approximately normal distribution.
      3. We can say that sales are approximately negatively skewed.
      4. It shows a negative kurtosis, the probability of normal distribution is higher than the extreme values.
   3. Plays-
      1. There are some advertisements with 0 plays with maximum plays as 63.
      2. Mean is approximately equal to median, hence it is a normal distribution.
      3. A density plot shows that Plays have imperfect normal distribution with flattened peak and thicker tails.
      4. Plays show negative Kurtosis.
   4. Attractiveness-
      1. Attractiveness is a factor with values ranging from 0 to 10.
      2. On an average, every play is rated with 6.77 rating. Best rated Play has 10 rating.
      3. Density plot shows that Attractiveness is negatively skewed.
      4. It has a positive Kurtosis that means the probability for normal distribution is higher than that for extreme values.
8. Exclude attractiveness for further analysis.
9. Check the correlation between Sales, Attractiveness and Plays-
   1. Advertise budges and sales are positively correlated, that means if we increase advertise budget, sales will go up.
   2. Advertisement budget and Plays are weakly positively correlated. Advertisement budget has very less impact on number of plays.
   3. Plays and Sales have a strong correlation. If we increase the plays, there is a high chance that the sales will also go up.
10. Now build a linear model using lm() function for sales against all other variables and assign it to a variable flm.
11. Examine fit by running this model.
    1. The intercepts for attractiveness is 11.3 which is not feasible.
    2. Attractiveness is a factor; hence use factor function to get a better model.
    3. There are 11 parameters including interceptor.
    4. Advertise and plays are affecting sales significantly.
    5. P value is 0.33 that is greater than 0.05, hence null hypothesis is rejected here.
    6. Multiplied R-squared value is 0.68 I.e., 68% of the predicted values can be correct.
12. Check the summary for this linear model using summary(flm) function-