Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Ans -

1. People prefer to rent bikes in summer and fall more than winter and spring

2. Casual customers prefer renting bikes on weekends/holidays whereas registered users seem to be using it on a daily basis for commute purposes may be

3. All users do not rent bikes in heavy rain situations

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

Ans - Any categorical variable can be explained by n-1 dummy values where n is the number of categories for the variable. So dropping first reduces the extra column and at the same time reduces the number of independent variables that we need to consider

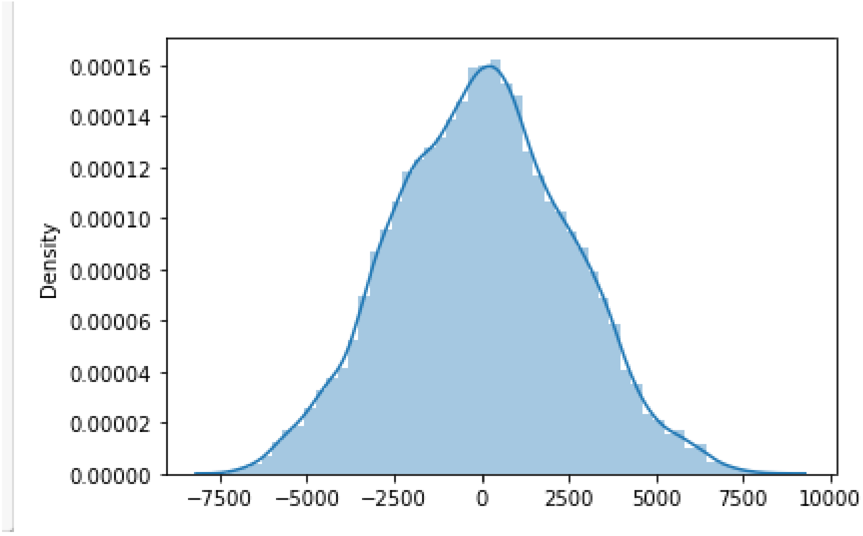
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans - Temp and atemp have the highest correlation with the target variable

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans – 1. Checked for VIF values for the independent variables to ensure minimal collinearity

2. The residuals follow normal or close to normal distribution for y_pred and y_test data



3. The residual errors are homoscedastic as in the variance does not increase towards either side of the distribution substantially

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans -

1. Temp/atemp
2. Year
3. Weathersituation

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

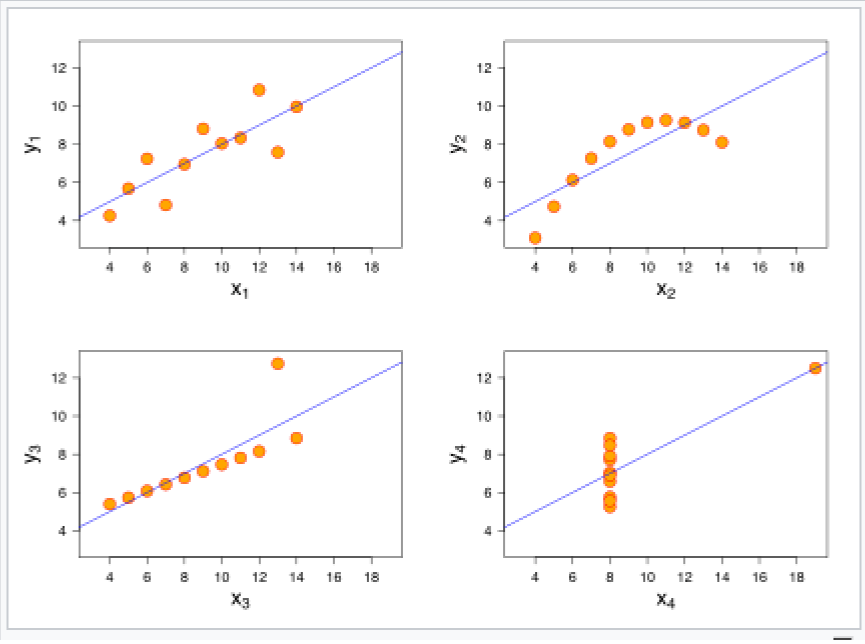
Ans – Linear regression algorithm tries to fit the dependent variable on independent variables in a linear fashion which can be explained by the equation – y = b0+b1x1+b2x2+b3x3…, where b1,b2,b3 are coefficients or slopes for the variables x1, x2 and x3 respectively and b0 is the intercept . There are following assumptions to a linear regression model –

1. **Linear functional form:** The response variable y should be a linearly related to the explanatory variables X.
2. **Residual errors should be i.i.d.**: After fitting the model on the training data set, the residual errors of the model should be independent and identically distributed random variables.
3. **Residual errors should be normally distributed**: The residual errors should be normally distributed.
4. **Residual errors should be homoscedastic**: The residual errors should have constant variance.

The null hypothesis for any independent variable's coefficient is B = 0 so if the p-value is under 5% or <0.05 then we can reject the null hypothesis and the variable becomes significant

2. Explain the Anscombe's quartet in detail. (3 marks)

Ans – Anscombe's quartet is a set of 4 datasets which look very similar when seen without any analysis or reference but follow very different distributions when plotted. It is used to demonstrate the value of visualizing data when analyzing it and the effect of outliers on the dataset



3. What is Pearson's R? (3 marks)

Ans - Pearson's R or Pearson's correlation coefficient is a measure of correlation between 2 variables. It is the ratio of covariance of 2 variables and product of their standard deviation. Correlation of 1 means as one variable increases, other does as well, correlation of 0 means the variables are unaffected by each other, correlation of -1 means as 1 variable increases, the other decreases. It's value can vary only between -1 and 1

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans – Scaling means to scale down or up the values to bring it closer to other variables so that the effect or coefficients are calculated on a more standard basis. Scaling does not effect predictions or accuracy, it only effects the values of the coefficient and helps visualize the data better. Normalized scaling or min max scaling is converting the entire data between 0 and 1 or -1 and 1, the scale is decided based on the max and min values of the variable so that the new values lie between 0 and 1. Whereas, standardized scaling changes the values to center them around the mean of the variable such that new mean becomes 0 and standard deviation becomes 1

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans – VIF is defined as 1/(1-R^2) so when R^2 is 1, VIF will become inf. It means that a variable is completely explained by other variables

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans – Q-Q plot or quantile-quantile plot is a plot of quantiles of 2 datasets.

Q-Q plots can be used to identify –

1. If 2 samples are similarly distributed or not based on the fit- line passing closely wrt to the plotted points or not
2. It can explain if the distribution scale is similar or not depending on the angle or slope of the fit-line
3. It can also be used to explain what kind of distribution best fits the sample data by fitting q-q plot between quantiles of the dataset and quantiles of different distribution(uniform/normal etc)