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)**

   From the dataset the there are some categorical variables that have a positive or negative effect on the dependent variable.

   Variables with Positive effect (Dependent variable value increases with increase in independent variables)

   a) Year

   b) Season (Summer, Winter)

   c) Month (September)

   Variables with Negative effect (Dependent variable value decreases with increase in independent variables)

   a) Holiday

   b) Weather Situation(Light Rain)

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

   Using drop_first=True will help reduce the extra column created during dummy variable creation. Hence it reduces the unnecessary correlations created among dummy variables.

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

   Looking at the pair-plot among the numerical variables the ‘temp’ has 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)**

   a) Linear Relationship: A scatter plot was plotted which showed linear relationship atleast for temperature.

   b) Multicollinearity: Validated that the VIF value of all the variables are within the acceptable range of 5 or less.

   c) Homoscedasticity: The variance is less at the beginning of the dataset and large thereafter and again decreased towards end. However there was no obvious pattern to conclude that it is not Homoscedastic.

   d) Error terms are normally distributed: A residual graph was plotted after the model which suggested that the error terms are normally distributed around zero.

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

   The top three features contributing significantly towards explaining the demand of the shared bikes are :

   a) temperature(0.5682)
   b) year (0.2334)
   c) lightrain (-0.2535)

General Subjective Questions

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

   Linear Regression Algorithm is a machine learning algorithm based on supervised learning. The steps to be followed to build a machine learning algorithm are mentioned below:

   a) Step 1: Reading, Understanding and Visualising Data

      Read the dataset and perform data cleaning and ETA steps.

   b) Step 2: Preparing the Data for modeling

      Add dummy variables, split the data into train and test and then do rescaling of the numerical variables.

   c) Step 3: Training the Model

      As the number of variables are large dropping one by one will be time consuming. So use RFE method to select an approximate number of variables on which you want to build the model. Look at the p value and VIF of the variables and decide which variable needs to be dropped.

      Please note the variable with high p value needs to be dropped first and then need to look at the VIF. Continue with building the models till all the variables have p-value less than 0.05 and VIF less than 5.

   d) Step 4: Residual Analysis of the Train Data

      Test for Homoscedasticity and normal distribution of error around 0.

   e) Step 5: Prediction and Analysis on Test Data

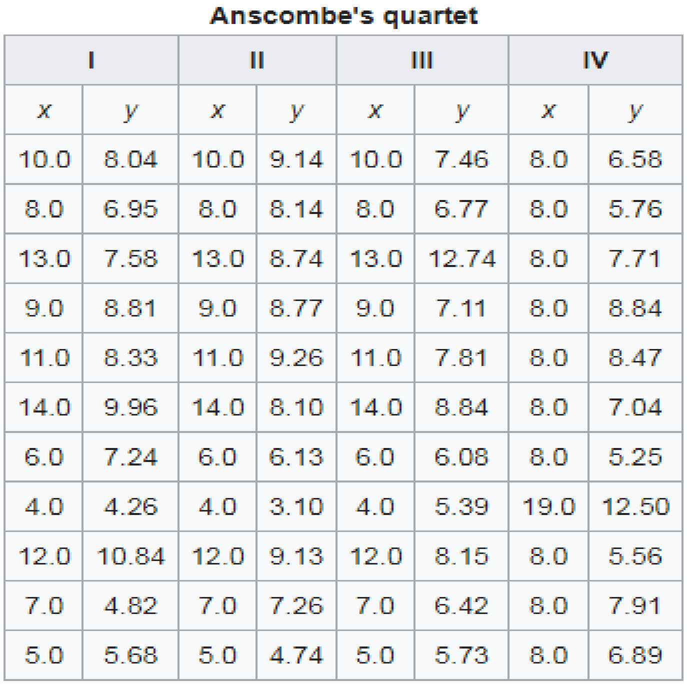
      Divide the dataset to X_test and y_test. Drop the columns which are not present in the train data set in the final model. Find out the r-square value of test data. If the difference between r-square value of test and train set is less than 5 % then the model can be considered a decent model.

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

   Anscombe's quartet is used to demonstrate the importance of graphical data when analyzing it, and the effect of outliers and other influential observations on statistical properties.

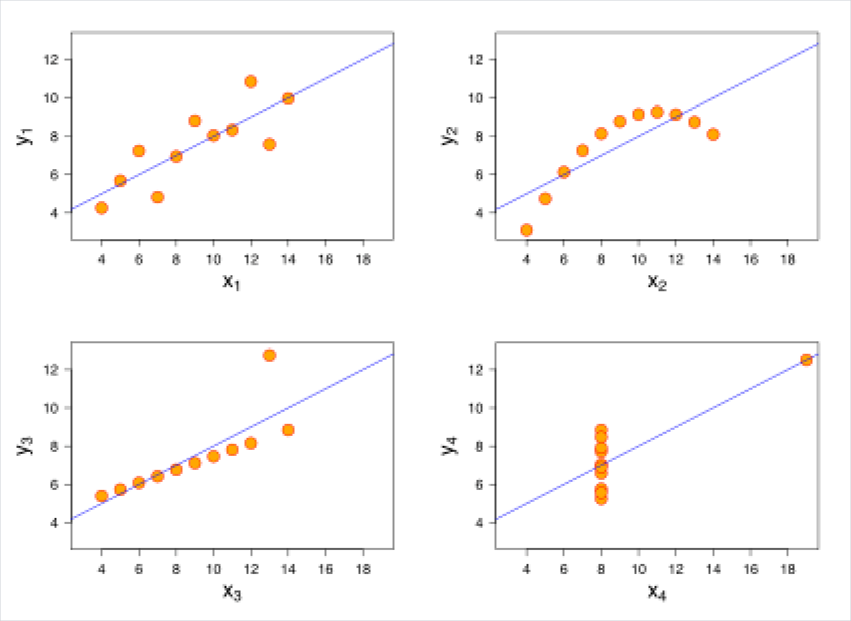
   It comprises of four data sets that have nearly similar statistics like mean, variance, correlation coefficient, line of best fit etc. However they have very different distributions and appear very different when graphed.

   The dataset and graph of Anscombe's quartet is provided below.



Anscombe's quartet

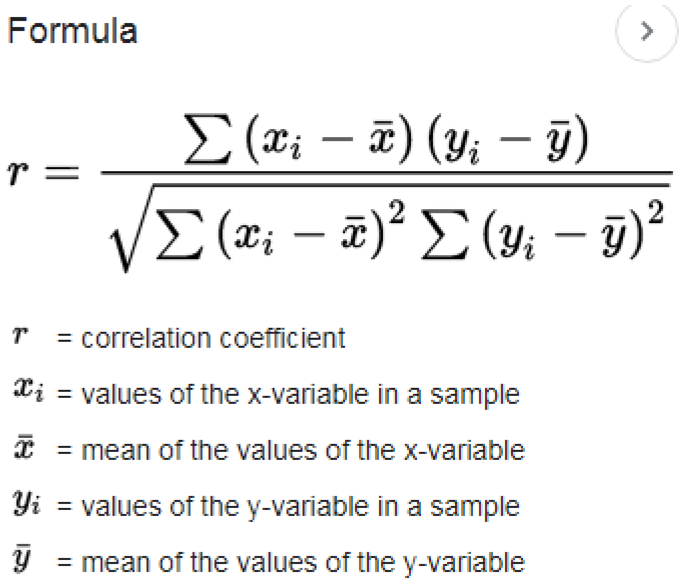
I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

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

Pearson's R is a correlation coefficient that is used in linear regression. It is used to measure the relationship between two variables. The R value can vary between − 1 and 1, where 0 is no correlation, 1 is strong positive correlation, and − 1 is strong negative correlation. For example if the correlation value between two variables is 0.8 it would mean that a significant and positive relationship exists between the two variables. A positive correlation indicates that if variable P goes up, then Q will also go up, whereas if the correlation value is negative, then if P increases, Q decreases. The formula for R is mentioned below:



Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

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

   Scaling is a step in pre-processing of data before fitting into the model. This is applied to the independent variables to normalize the data in a particular range for the numerical variables.

   Scaling is performed so that the variables have a comparable scale. If we do not have a comparable scale then some of the coefficients obtained by fitting the regression model would be very large or very small compared to other coefficients.

   Normalized scaling or Min-Max scaling is used to restrict the values between 0 and 1. However in standardized scaling the values are distributed in such a way that mean of the values is 0(zero) and the standard deviation is 1.

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

   VIF = 1/(1-R^2)

   So VIF can be infinite when R^2 value is 1. This means that a perfect fit of the model i.e. all movements of a dependent variable are completely explained by movements of the independent variable(s). This means that the model has over fitted.

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

   Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other.

Q-Q plot is used to fit a linear regression model, check if the points lie approximately on the line y = x. If the points do not lie on this line then residuals aren’t Gaussian and thus errors aren’t either. This implies that for small sample sizes, you can’t assume your estimator beta is Gaussian either, so the standard confidence intervals and significance tests are invalid.

A Q–Q plot is also used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.