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.

   There are 7 categorical variables in the dataset:
   - **year**: bike demand has substantially grown from 2018 to 2019
   - **holiday**: bike demand is less on holidays
   - **weather condition**: bike demand is more on clear days and goes down as the weather condition worsens
   - **season**: bike demand is high during summer & fall, and low during spring and winter
   - **month**: average no. of riders is high during may and oct; beyond which the demand gradually falls on both sides
   - **working day**: average no. of riders is similar on working days compared to non-working days
   - **day of week**: average no. of all riders shows similar patterns for every day of week

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

   Ans.

   It is important to use drop_first=True because it drops the first redundant dummy variables to prevent multicollinearity among the dummy variables.

   If a categorical variable with n-levels is used to create n nos. dummy variables, when n-1 variables have 0s, the variable left out will always be 1. So, value of any dummy variable can be inferred with the knowledge of remaining n-1 dummy variables and that makes the one of the n nos. dummy variables redundant. Keeping this redundant dummy variable can lead to multicollinearity among the dummy variables. Therefore, one of the dummy variables must be dropped.

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

   The variable **atemp** (feeling temperature in Celsius) has the highest correlation with the target variable in the pair plot of numerical variables as we can see the target variable moves closely with **atemp** (in the same direction, hence positive correlation) in an almost linear fashion and the data is not scattered haphazardly**.**

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

   The validation of assumptions was conducted as below:
   - **Linear relationship between target and at least one feature** – validated by creating scatter plots of (a) riders vs. temp (b) riders vs. wind speed and observing linear relationship in both plots
   - **Errors are normally distributed around zero mean** – validated by creating a distribution plot of the residuals and observing that it is very close to a normal distribution and centred around zero mean
   - **Homoscedasticity i.e., errors have a constant variance and do not have any pattern** – validated by plotting the residuals vs. target variable and observing that the it does not have any prominent pattern and nearly uniformly distributed on both sides of the zero line

- **Absence of Multicollinearity** – validated by plotting a heatmap of the correlation matrix of predictors of the final model and by observing that no two predictors are moderately/strongly correlated. The maximum correlation observed was 0.33, which is less than moderate.

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

   Top 3 Predictors are :

   temp_feels_like - temperature feels like positively affects the bike demand. Bike demand is likely to increase by 0.54 units for unit increase in temperature (temp_feels_like), provided other features remain unchanged
   weather condition - bad weather (snow or heavy snow) negatively affects the bike demand. Bike demand is likely to decrease by 0.28 units on snowy weather, provided other features remain unchanged
   year – increase in year positively affects the bike demand. Bike demand is likely to increase by 0.23 units every year, provided other features remain unchanged

General Subjective Questions

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

   Linear regression shows the relationship between two variables by applying a linear equation to observed data. One variable is supposed to be an independent variable, and the other is to be a dependent variable.

   In case of simple linear regression, data consisting of one independent variable(X) and one dependent variable (Y) is modelled as

   $Y = mX + c$
   where m is the slope of the line and c is the y-intercept (i.e. value of Y when X=0)

   The same model can be generalised for a multiple linear regression as

   $Y = B_0 + B_1X_1 + B_2X_2 + \ldots + B_nX_n$
   B_0 .. B_n are the coefficients and X_1 .. X_n are the independent variables

   The linear model's coefficients are obtained by minimizing the Sum of Square of Errors (Residual Sum of Squares or RSS) – this method is called Ordinary Least Square Method. The minimization of the cost function (RSS) can be done by using differentiation or by using gradient descent algorithm, which is an iterative process.

   The accuracy of the linear model is determined by the coefficient of determination or R^2, which is calculated as:
   $R^2 = 1 - (RSS/TSS)$

   RSS = Residual Sum of Squares = sum[$(y_i - y_{i\_pred})^2$]
   TSS = Total Sum of Squares = sum[$(y_i - y_{mean})^2$]

   The value of R^2 indicates what percentage of the variation in the data can be explained using the linear regression model.

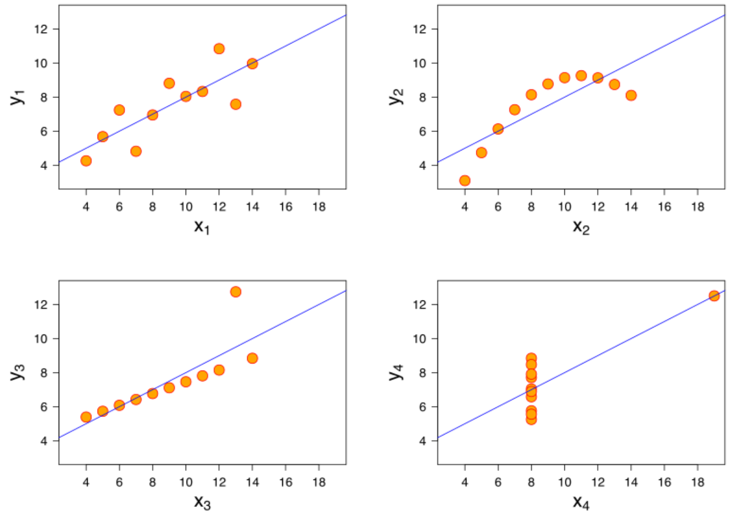
2. Explain the Anscombe's quartet in detail. (3 marks)

   Ans.

Anscombe's quartet is group of four datasets with nearly similar simple descriptive statistics viz. mean, variance, standard deviation, Pearson’s correlation coefficient, regression line, coefficient of determination etc., but have very different distributions and appear completely different when visualized. This group of 4 datasets2, each consisting of 11 nos. 2-variable (x, y) data points , was created by statistician Francis Anscombe to demonstrate the following:

- importance of data visualization when carrying out data analysis
- how outliers (extreme observations) affect the statistical properties

The plots of the quartets are depicted below:



3. What is Pearson’s R? (3 marks)

Ans.

Pearson’s R or Pearson's Correlation Coefficient is a statistic that measures the linear correlation between two variables. It has a numerical value that lies between -1.0 and +1.0.
The Pearson's Correlation Coefficient for a given a pair of random variables (X,Y) is calculated as

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

There are certain requirements for Pearson's Correlation Coefficient:

- Scale of measurement should be interval or ratio
- Variables should be approximately normally distributed
- The association should be linear
- There should be no outliers in the data

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

Feature Scaling is a method to bring the magnitude of the features of different ranges into a standard range. Feature scaling is done for (a) ease of interpretation, and (b) faster convergence of gradient descent.

Feature scaling does not change the model accuracy or the statistical significance of the coefficients. However, it changes the coefficient values.

Normalized scaling or min-max scaling results in scaling the features between 0 and 1. The formula for the same is
$X_{normalized} = (X-X_{min})/(X_{max}-X_{min})$

Standardized (Z-score) scaling method scales the features in a way that the scaled feature has mean=0 and standard deviation=1. The formula for the same is

$X_{standardized}$ = (X-mean(X))/Std. Dev(X)

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

Ans.

If VIF of a predictor is infinity, it means that the predictor has a perfect linear relationship with the rest of the predictors. The formula for VIF is

$$\text{VIF} = \frac{1}{1 - R^2}$$

When the predictor has a perfect linear relationship with the rest of the predictors, the coefficient of determination (R^2) becomes one and VIF becomes infinity.

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

Quantile-Quantile or Q-Q plot is a graphical method for determining whether two samples of data came from the same population or not. Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Quantile means the fraction (or percent) of data points below the given value.

If we want to check if a sample is normally distributed or not, then we calculate the quantile values of the sample data points and the theoretical quantile values for a normal distribution and create a scatter plot. Theoretical quantile values of a normal distribution is obtained by dividing the normal distribution into same nos. of equal probability zones (areas). If the scatter plot follows a straight line, we can conclude that the sample data is normally distributed.

In case of linear regression, one of the assumptions is normal distribution of error terms (residuals). By plotting the Q-Q plot with y-values as the quantile values of residuals and x-values as the corresponding quantiles of a standard normal distribution, we can observe if the relationship is near linear or not. A linear relationship will validate the assumption of normally distributed error terms with zero mean.