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?**

A. **Season**:
The demand for bikes is less in the month of spring when compared to other seasons.
Year:
The demand for bikes increased in the year 2019 than in 2018.
Month:
The demand for bikes is high from June to September and It is also observed that January has the lowest demand.
Holiday:
The demand for bikes is less in holidays in comparison to not being a holiday.
Weekdays:
The demand for bikes is almost similar throughout the weekdays.
Workingday:
There is no significant change in bike demand with working days and non working days.
Weathersit:
The bike demand is high when the weather is clear and Few clouds. But, the demand is less in case of Light snow and light rainfall. We do not have any data for Heavy Rain + Ice Pellets + Thunderstorm + Mist, Snow + Fog. So, we cannot reach any conclusion. Maybe the company is not operating on those days or there is no demand for bikes in that weathersit.

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

A. pd.get_dummies() creates 'n' dummy variables for a categorical variable with 'n' levels. But we need only 'n-1' dummy variables to represent 'n' dummy variables. So we use 'drop_first=True' for dummy variable creation.

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

A. Based on the analysis, it is observed that both 'atemp' & 'temp' variables are highly correlated with the target variable 'cnt'. Both have the correlation value of 0.63 with 'cnt'.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

A.

Assumption 1:
The Dependent variable and Independent variable must have a linear relationship.
Here, I have plotted pair plot between independent & dependent variables to show that the independent variables exhibit linear relationship with dependent variable (target variable)

Assumption 2:
No Autocorrelation in residuals.
Here, I have taken the value of Durbin-Watson - 2.016 from lm.summary() which states that there is no autocorrelation.

Assumption 3:
No Heteroskedasticity
Here, I have plotted the scatter plot on Residual vs Fitted values and observed that the plot is not in a funnel shape. Hence, No Heteroskedasticity

Assumption 4:
No Perfect Multicollinearity.
Here, I have calculated VIF values of all variables and I got all the VIF values of variables < 3 which clearly implies that this assumption is correct.

Assumption 5:
Residuals must be normally distributed with mean as 0.
Here, I have plotted a distplot on residuals and checked that the plot has normal distribution with mean as 0.

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

A. The top features that contributes significantly towards bike demands w.r.t model are:

- Feel like temperature (atemp) with coefficient value +0.467
- Humidity (hum) with coefficient value -0.28
- Year (yr) with coefficient value +0.231

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

A. Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. When the number of the independent feature is 1 then it is known as Univariate Linear regression or Simple Linear regression, and in the case of more than one feature, it is known as multivariate linear regression or Multilinear regression.

The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables. The equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variables.

$$Y = \beta 0 + \beta 1X1+\beta 2X2+...+ \beta pXp+\varepsilon$$

Y - a dependent or target variable
$X1,X2...Xp$ - independent variables also known as the predictors of Y
Interpretation of coefficients(β1,β2....):
Change in mean response E(y), per unit increase in the variable when other predictors are held constant

Assumption for Linear Regression Model

Linear regression is a powerful tool for understanding and predicting the behavior of a variable, however, it needs to meet a few conditions in order to be accurate and dependable solutions.

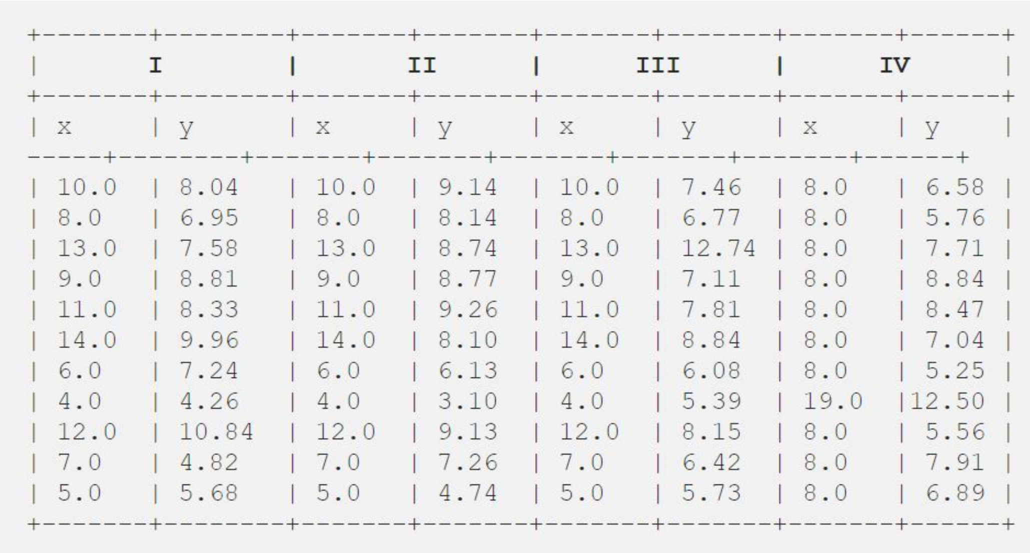
1. **Linearity**: The independent and dependent variables have a linear relationship with one another. This implies that changes in the dependent variable follow those in the independent variable(s) in a linear fashion.
2. **Independence**: The observations in the dataset are independent of each other. This means that the value of the dependent variable for one observation does not depend on the value of the dependent variable for another observation.
3. **Homoscedasticity**: Across all levels of the independent variable(s), the variance of the errors is constant. This indicates that the amount of the independent variable(s) has no impact on the variance of the errors.
4. **Normality**: The errors in the model are normally distributed.
5. **No multicollinearity**: There is no high correlation between the independent variables. This indicates that there is little or no correlation between the independent variables.

2. **Explain the Anscombe's quartet in detail.**

A. Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

Anscombe's quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph.

The four datasets of Anscombe's quartet are



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

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

3. **What is Pearson's R?**

A. The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Pearson correlation coefficient (r)	**Correlation type**	**Interpretation**
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the same direction.
0	No correlation	There is no relationship between the variables.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the opposite direction.

The Pearson correlation coefficient is a good choice when all of the following are true:

1. Both variables are quantitative
2. The variables are normally distributed
3. The data have no outliers
4. The relationship is linear

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

A. It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the time, the collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then the algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like **t-statistic, F-statistic, p-values, R-squared**, etc.

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and 1.
- *sklearn.preprocessing.MinMaxScaler* helps to implement normalization in python.

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

- One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$x = \frac{x - mean(x)}{sd(x)}$$

- *sklearn.preprocessing.scale* helps to implement standardization in python.

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

A. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.
A general rule of thumb is that if VIF > 10 then there is multicollinearity. Note that this is a rough rule of thumb, in some cases we might choose to live with high VIF values if it does not affect our model results such as when we are fitting a quadratic or cubic model or depending on the sample size a large value of VIF may not necessarily indicate a poor model.

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

A. Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.

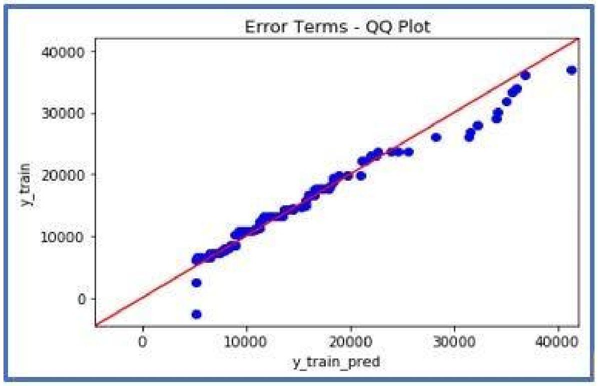
It helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with the same distributions.

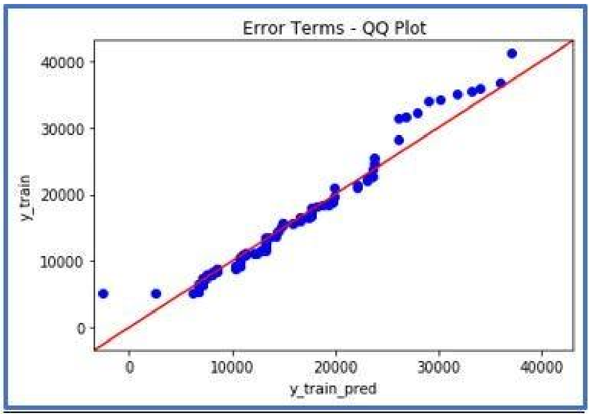
Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Below are the possible interpretations for two data sets.

1. **Similar distribution:** If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
2. **Y-values < X-values:** If y-quantiles are lower than the x-quantiles.



3. **X-values < Y-values:** If x-quantiles are lower than the y-quantiles.



4. **Different distribution:** If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis