## Assignment-based Subjective Questions

**Q-1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

**Ans:** Following conclusions were drawn for respective variables:

**Yr (year)-** There was a significantly higher sales happened in the year 2019 with comparatively very high median.

**Season-** People tend to rent bikes in significantly very higher number in fall season, whereas the bike rentals recorded in spring was comparatively very less.

**Holiday-** Very high variance was in Holiday in bike rental counts

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

**Ans:** We just want to create **n-1 dummy** variables for n unique values of **each categorical variable**, that's why we drop the first dummy variable from the set of n variables.

Let's understand why don't we need all n variables:

**For example**, in our Problem statement we have "season" as the categorical variable which has more than 2 unique values namely- fall, winter, summer and spring. Suppose while creating the dummy variables we drop the "fall" variable using drop_first, then whenever there will be 0 value for rest of the three (winter, summer, spring) that will indicate the value 1 for "fall" variable, that's why we just need only n-1 variables in the dataset to avoid redundancy.

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

**Ans:** "**temp**" and "**atemp**" variables had the highest correlation with target variable "count" which was equal to **0.63**.

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

**Ans: Independent Error terms:** Validated the independence of the error terms by plotting the scatter plots between residuals and independent variables in dataset.

**Homoscedasticity:** The variance is approximately constant through the scatter plots.

**Normality of errors:** Visualized a normal curve by plotting histogram of residuals.

**Multicollinearity:** Eliminated the variables with very high VIF values to remove multicollinearity.

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

**Ans:** The top 3 features with respective coef. Values are temp = 0.55, hum= -0.32, windspeed= -0.23

## General Subjective Questions

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

**Ans:** When we want to predict the target variable having **continuous values** using all independent variables **(with any datatype)**, in that case we use regression machine learning technique. It's a **supervised learning** algorithm, meaning it learns from existing labeled data to predict future outcomes.

Now if want to **draw a linear relationship** between the independent variables and the continuous target variable, then we use linear regression algorithm.

This algorithm fits the line between the independent and dependent variable on x, y plane in case of **Simple linear Regression** (when data has only one independent variable)

Whereas in **Multiple linear Regression** a hyperplane will be predicted and fit between the independent variable and dependent variable.

**Model representation for linear regression:**

Y = B0 + B1X1 + B2X2 + B3X3 + .... + BNXN

where,

Y - dependent or target variable

N - number of independent variables

X1, X2, X3 – independent varaibles

B1, B2, B3 – coefficient for corresponding independent variables

B0 – intercept

**Finding the Best-Fitting Line:**

Now to find the best values for the coefficients (B1, B2, B3 and so on) or strength of relationships between each independent variable and dependent variable we utilize the methods such as

Ordinary least squares or Gradient descent to minimize the squared distance between linear regression line and the actual values of dependent variable. This process is called **Parameter estimation** or **least squares estimation.**

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

**Ans:** It is set of 4 datasets to illustrate the importance of **data visualization** in statistical analysis, particularly for linear regression.

**Components of Anscombe's Quartet:**

Each dataset consists of 11 data points with one independent variable (x) and one dependent variable (y). The key point is that all four sets share the following characteristics in terms of their calculated summary statistics:

- **Equal Mean (x):** All sets have an average x-value of 9.
- **Equal Mean (y):** All sets have an average y-value of 7.5.
- **Equal Variance (x):** The spread of x-values is identical across all sets.
- **Equal Variance (y):** The spread of y-values is identical across all sets.
- **Equal Correlation Coefficient:** The linear relationship between x and y appears similar based on the correlation coefficient.
  pen spark

**Importance of Visualization:**

While the summary statistics suggest a similar underlying relationship between x and y in all four datasets, creating scatter plots reveals a completely different story:

- **Dataset 1:** Shows a clear linear relationship between x and y.
- **Dataset 2:** Exhibits a seemingly random scatter with no discernible pattern.
- **Dataset 3:** Reveals a strong non-linear relationship, with a tight cluster of points followed by a single outlier.
- **Dataset 4:** Has a curved relationship, with an outlier pulling the line in a misleading direction.

**Q.3: What is Pearson's R?**
**Ans**. Pearson's correlation coefficient used to measure the correlation between the two variables, it represents the magnitude as well as the direction of the relationship between the two variables. the values could range between (-1 to 1), if it's -ve that means if one increase other will decrease, if it's +ve that means if one increase other will also increase.

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

Ans: Scaling is the technique to bring all the variable of the dataset to the same range of numerical values.
Scaling doesn't impact the accuracy of the model, but it can speed up the model building process as all the variables are in the same numerical range and also it makes the gradient descent to converge faster.

**Standardized Scaling:**

- It centers the data around mean of 0 having standard deviation of 1
- In case of normal distribution 99.7% values will fall under -3 to 3 as it has SD equal to 1.
- It is calculated as z = (x - mean) / standard deviation.

**Normalized scaling:**

- The values will be between -1 to 1 or 0 to 1 depeding on the scaling we choose.
- Here's a formula in case if we use min max scaling normalized value = (x - minimum value) / (maximum value - minimum value).

**Q.5 You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
**Ans:** This can happen in case when an independent variable can be completely defined by the linear combination of other variables.
One more case could be there is a very solid correlation between two independent variables (perfect collinearity). And according to, VIF = 1/(1-R^2), when R^2 becomes 1, the VIF will tend to infinity.

**Q.6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
**Ans:** In linear regression it is used to compare the quantiles of two distributions, and it is vital technique for assessing an assumption of linear regression which is **Normality of Residuals.**

**Importance of Q-Q Plots in Linear Regression:**

- **Normality Assumption:** Linear regression often relies on the assumption that the residuals (errors) are normally distributed. This assumption is crucial for the validity of hypothesis tests and confidence intervals associated with the model coefficients.
- **Identifying Issues:** Q-Q plots provide a visual way to assess this normality assumption. Deviations from the straight line can warn you about potential problems with the model or the data.
- **Alternative Approaches:** If the Q-Q plot reveals non-normal residuals, you might need to consider alternative regression techniques like robust regression or transformations of the data to achieve normality.
- **Outlier Detection:** Q-Q plots can also help identify outliers in the residuals, which can affect the model's performance. Points far from the diagonal line might indicate outliers that deserve further investigation.
- **Model Comparison:** You can use Q-Q plots to compare the normality of residuals from different linear regression models and choose the one with the closest fit to a normal distribution.