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?

Ans :

- **Seasons**: Summer and winter positively influence bike rentals, while spring negatively influences them.
- **Months**: September sees increased rentals, whereas July sees a decrease.
- **Weather Conditions**: Adverse weather conditions (light snow/rain and misty/cloudy) negatively impact bike rentals, with light snow/rain having a more substantial effect.
- **Yearly Trend**: There is a strong positive trend over time, indicating increasing bike rentals.
- **Holidays**: Bike rentals decrease on holidays.

These insights suggest that bike-sharing companies should tailor their strategies to account for seasonal variations, weather conditions, and temporal trends to optimize bike usage and enhance service delivery.

For eg, Do analysis on what we can do to increase the usage of bike rentals in adverse weather conditions, It would be to provide appropriate clothing/equipment to drive safely in light rains/snow? Or is it because people generally are not going out on those days?

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

Ans:

The drop_first=True argument is used to avoid the scenario in which the independent variables are multicollinear - a scenario in which two or more variables are highly correlated; in simple terms one variable can be predicted from the others.

For eg consider weekday in our case, if its not a Mon, Tue, Wed, Thur, Friday or Saturday, for sure it is Sunday. By eliminating this Sunday from the list, we are eliminating “perfect” multicollinearity.

In a multiple regression model, each coefficient theoretically represents the change in the response variable for each unit change in that predictor, assuming all other predictors are held constant. However, when predictors are highly correlated (multicollinearity), it means they change together, so it’s difficult to tease apart the effect of individual predictors on the response variable.

While multicollinearity doesn’t affect the model’s ability to predict the response variable, it does affect our understanding of the individual predictors.

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

Feeling temperature (.65) and Temperature in Celsius (.64) has highest correlation with target variable.

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

1) **Residual analysis** - check if the error terms are also normally distributed which is infact, one of the major assumptions of linear regression

2) **Homoscedasticity** - This assumption means that the variance of the errors is constant across all levels of the independent variables. It can be checked by plotting residuals vs. predicted values. The plot should show a random pattern of points with a roughly constant variance.

3) **Lack of Multicollinearity**: Multicollinearity occurs when the independent variables are too highly correlated with each other. Multicollinearity can be detected using Variance Inflation Factor (VIF) or the correlation matrix. A VIF value of more than 5 is generally considered as having high multicollinearity.

4) **Cross-Validation**: The model is trained on the training set and then the predictive accuracy is assessed on the validation set.

5) **Performance Metrics :** Root Mean Squared Error (RMSE), R-squared, F-stats etc are within limits

6) **Observed vs. predicted values** – In a well-fitting regression model, the points should lie close indicating that the predicted values are close to the observed values.

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

1) Temperature (coef : .49, t-value: 14.8, p value: 0) – Strong positive impact, Higher temperatures lead to increased bike usage.
2) Year (coef : . 0.2335, t-value: 28.361, p value: 0) - Strong positive impact, a strong upward trend in bike rentals over the years.
3) Weather Situation – Light Snow/Rain (coef : -0.2852, t-value: 14.8, p value: 0) - Strong negative impact - Adverse weather conditions significantly decrease the number of bike rentals, highlighting the sensitivity of bike usage to weather changes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

A linear regression model attempts to explain the "linear" relationship between a dependent and independent variables. Simple linear regression has only one independent variable while multiple linear regression has multiple independent variables on which the dependent variable is dependent on.

The mathematical representation can be shown as : y = beta0 + beta1*x1 + ... + beta(n) * x(n) + e
Where x1, x2 are features and beta0 –beta(n) are weights.

Our aim is to find the optimal weights for these features (independent variables) from the observed values, so that we can use the equation to predict the dependent variable y given we know values of independent variables.

The coefficients are estimated using a method called least squares which minimizes the sum of squared residuals (differences between the observed and predicted values).

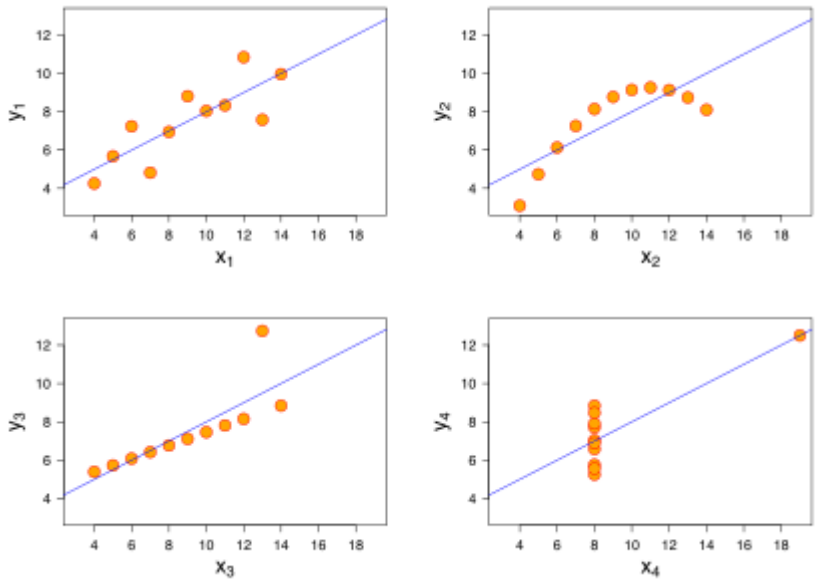
Gradient Descent is an optimisation algorithm which optimises the objective function (for linear regression it's cost function) to reach to the optimal solution. It adjusts the parameters by moving in the direction of the negative gradient to decrease the function.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises a set of four datasets, having i**dentical** descriptive statistical properties in terms of **means, variance, R-squared, correlations, and linear regression lines** but having different representations when we scatter plots on a graph.

They were constructed by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the **effect of outliers and other influential observations** on statistical properties.

It gives an amazing illustration of why not to depend on summary statistics and why **visualization is important**!



3. What is Pearson's R?

Pearson's R, is a statistical measure that determines the strength and direction of the linear relationship between two quantitative variables.

The formula for Pearson's R for population is :
Pearson's R = product of standard deviations/covariance of the two variables.

The value of Pearson's R ranges from -1 to +1:

- A value of -1 indicates a perfect negative linear correlation, meaning as one variable increases, the other decreases.
- A value of +1 indicates a perfect positive linear correlation, meaning both variables increase or decrease together.
- A value of 0 indicates no linear correlation between the variables.

It's important to note that Pearson's R only measures linear correlations and may not useful if it's nonlinear.

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

Scaling adjusts the spread or variability of your data, Scaling is a method used to standardize the range of independent variables or features of data.
It's a crucial step in pre-processing your data, especially when the dataset contains variables of different scales.
Without scaling,
a) The model could become biased or skewed towards the variables with higher magnitude.
b) Scaling will make models less sensitive to outliers.

For eg : lets consider a model evaluating house price from area and bedrooms. Model tends to provide more weightage to area since its in thousands and bedrooms are in single digit numbers.

Difference between Normalized Scaling and Standardized Scaling:
Normalized Scaling (Min-Max Scaling): Rescales the data to a specific range, usually between 0 and 1.
The new value is calculated as: Xnew=Xmax−Xmin/X−Xmin. Normalization is useful when there are no outliers as it cannot cope up with them.
It is often used when features are of different scales
Standardized Scaling (Z-Score Normalization): This technique transforms data to have a mean of 0 and a standard deviation of 1. The new value is calculated as: Xnew=X−μ /sigma
where μ is the mean and sigma is the standard deviation. Standardization can be helpful in cases where the data follows a Gaussian distribution.
It is not affected by outliers because there is no predefined range of transformed features.

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

If the VIF is infinite, it usually means that one of predictor variables (features) is a ***perfect*** linear combination of one or more of the other features. In other words, one variable can be exactly predicted from the others. This is a situation of perfect multicollinearity. For eg - feature1 = feature2 * 2

We need to transform the dependent features (combine etc) or drop the features to get rid of infinite VIF.

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

A Q-Q plot, short for Quantile-Quantile plot, is a graphical way to assess if a set of data came from some form of theoretical distribution such as a Normal, Exponential, or Uniform distribution.

For example, you might collect some data and wonder if it is normally distributed. A QQ plot will help you answer that question.

You can also use QQ plots to compare to different datasets that you collected to determine if their distributions are comparable.

In linear regression, a Q-Q plot is used to check the normality of the residuals.

The residuals of a linear regression model are the differences between the observed and predicted values of the dependent variable.

If these residuals are normally distributed, it validates one of the key assumptions of linear regression.

Here's how a Q-Q plot works:

The quantiles of the dataset are plotted against the quantiles of a theoretical distribution, usually the standard normal distribution.

If the points in the Q-Q plot lie approximately on a straight diagonal line, it suggests that the data is normally distributed.