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)

Answer: Amongst the 5 categorical variables, here are the key inferences:

a) Working Day – Demand for bikes is more during working day as they are mostly used for commuting to work.
b) Weather – Demand for bike rentals is more when the weather is clear.
c) Season – Summer and Winter are linked to high demand.
d) Month – June to September is the peak season for bike rentals.
e) Weekday – Not having any influence on demand

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

Answer: drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation and hence avoid multicollinearity. Thereby, it reduces the 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)

Answer: 'temp' variable 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)

Answer: The assumptions of Linear Regression were validated through the following steps:

1. **Linearity**: Checked by plotting the residuals against the predicted values; a random scatter suggests linearity.
2. **Homoscedasticity**: Ensured by observing the residuals; a consistent spread across the residual plot indicates homoscedasticity.
3. **Normality of Residuals**: Verified using a Q-Q plot; residuals should follow a straight line if they are normally distributed.
4. **Multicollinearity**: Addressed by calculating the Variance Inflation Factor (VIF); predictors with VIF > 5 were considered for removal.
5. **Independence of Errors**: Durbin-Watson statistic was used to check for autocorrelation in residuals; a value close to 2 suggests independence.

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

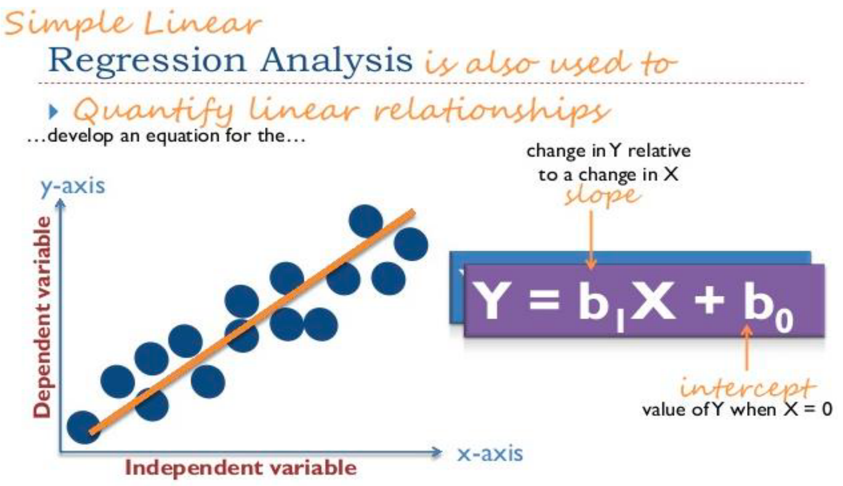
<u>Answer:</u> Based on the final model, the top 3 features contributing significantly are : temperature, weather situation and year.

General Subjective Questions

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

Answer:

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.



The purpose of regression analysis is calculate estimates of the slope and intercept.

While training the model we are given:
X: input training data (univariate –one input variable(parameter))
Y: labels to data (supervised learning) When training the model –it fits the best line to predict the value of Y for a given value of X. The model gets the best regression fit line by finding the best b1 and b0 values.
b0: intercept
b1: coefficient of x

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

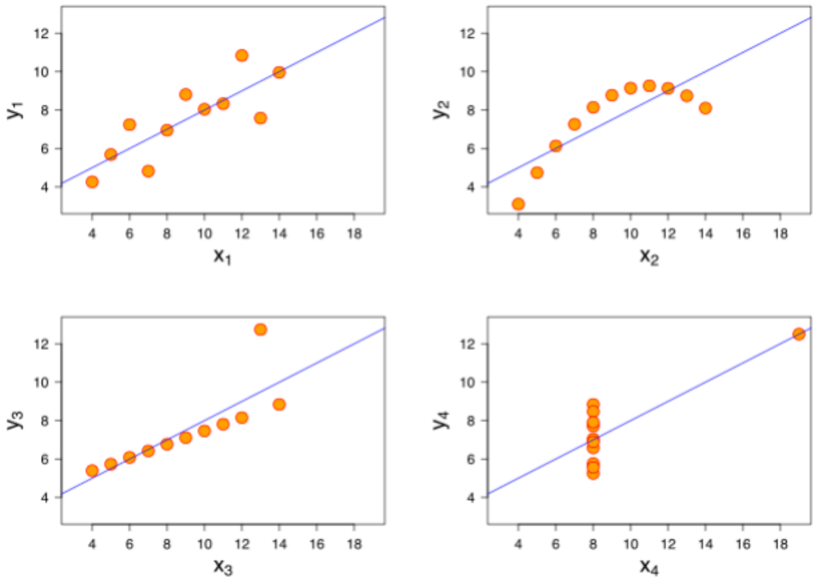
Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.

- The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.

- In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.



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

Correlation between sets of data is a measure of how well they are related. The most common measure of correlation in stats is the Pearson Correlation. The full name is the Pearson Product Moment Correlation (PPMC). It shows the linear relationship between two sets of data. In simple terms, it answers the question, Can I draw a line graph to represent the data? Two letters are used to represent the Pearson correlation: Greek letter rho (ρ) for a population and the letter “r” for a sample.

Potential problems with Pearson correlation

The PPMC is not able to tell the difference between dependent variables and independent variables. For example, if you are trying to find the correlation between a high calorie diet and diabetes, you might find a high correlation of .8. However, you could also get the same result with the variables switched around. In other words, you could say that diabetes causes a high calorie diet. That obviously makes no sense. Therefore, as a researcher you have to be aware of the data you are plugging in.

In addition, the PPMC will not give you any information about the slope of the line; it only tells you whether there is a relationship

Real Life Example

Pearson correlation is used in thousands of real life situations. For example, scientists in China wanted to know if there was a relationship between how weedy rice populations are different genetically. The goal was to find out the evolutionary potential of the rice. Pearson’s correlation between the two groups was analyzed. It showed a positive Pearson Product Moment correlation of between 0.783 and 0.895 for weedy rice populations. This figure is quite high, which suggested a fairly strong relationship.

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

What is scaling? Why is scaling performed?

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range It is performed during the data pre processing to handle highly varying magnitudes or values or units If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values

Difference between normal standardized scaling

Normalization is a good technique to use when you do not know the distribution of your data or when you know the distribution is not Gaussian (a bell curve) Normalization is useful when your data has varying scales and the algorithm you are using does not make assumptions about the distribution of your data, such as k nearest neighbors and artificial neural networks

Standardization assumes that your data has a Gaussian (bell curve) distribution This does not strictly have to be true, but the technique is more effective if your attribute distribution is Gaussian Standardization is useful when your data has varying scales and the algorithm you are using does make assumptions about your data having a Gaussian distribution, such as linear regression, logistic regression, and linear discriminant analysis.

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

What is VIF
The variance inflation factor VIF quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing collinearity/ multicollinearity Higher values signify that it is difficult to assess accurately the contribution of predictors to a model.

Why is VIF infinite
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. It’s just a visual check, not an air tight proof, so it is somewhat subjective. But it allows us to see at a glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that’s roughly straight. Here’s an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.