Bike Sharing – Linear Regression Assignment

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)**
   A: The biggest takeaways from the categorical variables are:
   a. Summer and winter are the best months for the company.
   b. A weather situation of Light snow, misty+cloudy negatively affect the bike rentals
   c. More bikes are rented on working days as compared to holidays.

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   A: As a rule, if we have n values in a categorical variable, we only need **n-1** dummy columns to accurately represent a data. For e.g., a categorical variable with just 2 values would need just one dummy column (with a 0 or 1) to accurately represent the data. For this reason, we can drop the first column to still accurately represent the data and reduce redundant data and avoid multi collinearity.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**
   A: **Registered** has the highest correlation with count followed by **casual**. But since these two values together make up count, if we were to ignore the two columns, **temperature** has very high correlation with count.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
   A: This most important assumption is that the linear regression needs the relationship between the independent and dependent variables to be linear. We validated this assumption by plotting a **scatterplot** and confirming that a lot of features did indeed show linear relationship.

   The second assumption assumes that the error terms are distributed normally and are centred around zero. We validated this assumption using a **histogram** in the project.

   Another assumption is that of homoscedasticity. It basically means residuals have constant vadiance at every level of x. This assumption was validated in the project by plotting **residual vs fitted values**.

   Finally, linear regression also assumes minimal multicollinearity and we ensured this was the case by **verifying VIFs** of the features in the model.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**
   A: Temparature, weathersit and humidity

General Subjective Questions

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

   A: Linear Regression is a machine learning algorithm based on supervised learning. Supervised learning involves training an algorithm based on labelled or training data. Linear regression models commonly used for modelling the relationship between a single dependent variable(y) and one or more predictors(X). When we have one predictor, we call this "simple" linear regression:

   $$E[Y] = \beta 0 + \beta 1X$$

   That is, the expected value of Y is a straight-line function of X. The betas are selected by choosing the line that minimizing the squared distance between each Y value and the line of best fit. The betas are chosen such that they minimize this expression:

   $$\sum i\ (yi - (\beta 0 + \beta 1X))2$$

   There are certain assumptions made while using linear regression models which are critical. These are:

   a. Linearity: The relationship between X and the mean of Y is linear.
   b. Homoscedasticity: The variance of residual is the same for any value of X.
   c. Independence: Observations are independent of each other.
   d. Normality: For any fixed value of X, Y is normally distributed.

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

   A: Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyse it and build your model. Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics. However, if one were to plot these datasets on a graph, you will notice that they have very different distributions and appear completely different.

   The four datasets of Anscombe's quartet are as below:

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

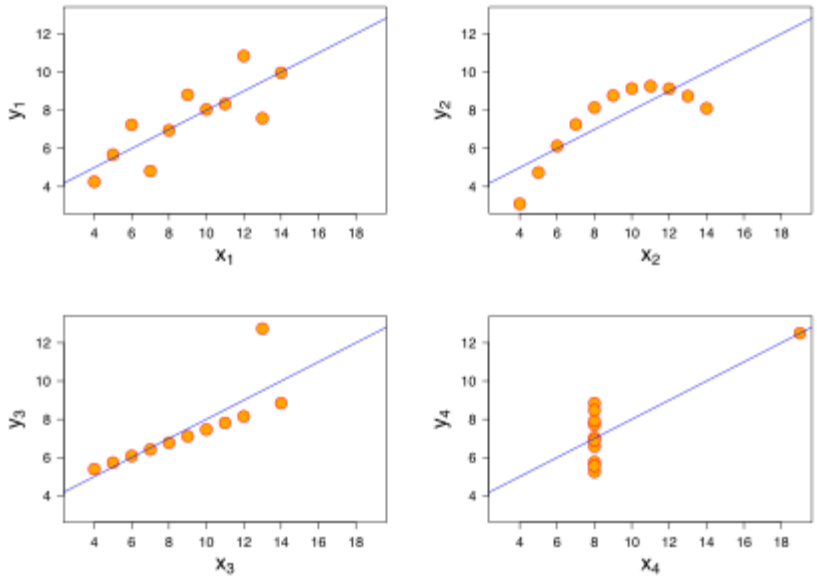
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Please take note that:

- Mean of x is 9 for all four datasets.
- Sample variance of x is 11 for all four datasets
- Mean of y is 7.50 for all four datasets
- Sample variance of y Is 4.125 for all four datasets
- Correlation between x and y is 0.816 for all four datasets.
- Linear regression line is y=3+0.5x for all four datasets
- Coefficient of determination of linear regression R^2 is 0.67 for all four datasets.

From above, you can see that all four datasets have identical simple descriptive statistics.

However, when we plot the datapoints as a scatter plot, we see below:



The first scatter plot (top left) appears to be a simple linear relationship The second graph (top right); while a relationship between the two variables is not linear.In the third graph (bottom left), the modelled relationship 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.

The above datasets are a perfect example of why looking at a set of data graphically is critical to analyze data and its patterns.

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

   A: Pearson's R is a measure of linear correlation between two sets of data. It is the ration between the covariance of two variables and the product of standard deviations. It is a normalized measurement of covariance. Hence it lies in the range of -1 and 1. The meaning of Pearson's R can be summarized as below:

   - r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
   - r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
   - r = 0 means there is no linear association

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

   Scaling is a geometric change that linearly enlarges or reduces things. It is a critical step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range.

   Different features in machine learning may have different magnitudes, units and range and scaling can be used to bring all the features in the same standing so that we can increase readability and understand the impact features has on a target better. If scaling is not done,

algorithm may only take magnitude into account and not units and hence may result in an incorrect model.
Differences between normalisation and standardization can be summarized as below:

NO.	Normalisation	Standardisation
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.

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

A: VIF stands for Variance inflation factor. In VIF, each feature is a regression against all other features. The formula for VIF is:

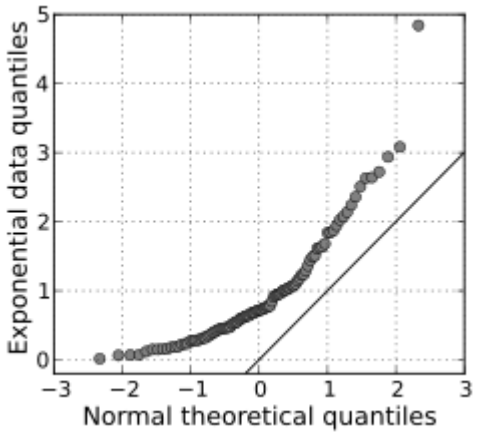
$$VIF = 1 / (1 – R2)$$

R2 increases as correlation between the feature and other features increases, if there is perfect correlation, that would make R2 as 1 and VIF becomes infinite. To reduce VIF in this case(or in general with a high VIF), we need to drop one of the variables from the dataset which is causing the high multicollinearity.

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

A: Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. A goof Q-Q plot of Residuals vs theoretical quantiles has all of the residuals lying on or close to the line.