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?

Below are the categorical variables found in the dataset and there effect on dependent variable “cnt”

1. Season : We observed, cnt is lowest in season of “spring”, increases modrately in summer, increase to highest during fall and then dip during winter.

2. yr : We observed, there is significant increase in bike sharing “cnt” from 2018 to 2019. Hence every year there will be an increasing trend as per the data. But we can not confirm on this with only 2 years data.

3. mnth : We observed, every year January remains the Lowest bike sharing Month, gradual increase in bike sharing from Feb to June. After that bike sharing remains at peak during months of June, July, Aug and Sep and the gradually dips down again till december and January

4. holiday : We observe, there is significantly large fluctuation in Min and Max cnt during Holiday than in non-holidays. However we observed there is no significant diffference between the cnt value during Holidays and Non-Holidays.

5. Weekday : No significant variations observed between the Weekday.

6. Workingday : we observed there is no significant diffference between the cnt value during Working-day and Non-Workingday.

7. Weathersit : Weathersit is having the most significant pattern to observe. The Bike sharing is at peak when the sky is clear or with less cloud. There is a significant dip in Bike sharing observed if the day is Cloudy & Misty. There is further more dip in Bike sharing if during Light Snow and Rain. And then the Least during

Heavy rain and Snow. Hence the weather plays a very significant role in Bike sharing business.

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

   drop_first=True drops the redundant column which may simply add up to number of column. For example if we have categorical variable gender which is created a dummy variable Male and Female. Here Male will have two values 1 and 0. 1 representing Male and 0 representing Not Male or we can also say 0 = Female. Hence even if we drop one dummy variable female. We still have all the information intact.
   Necessity to drop one column is, to avoid Multi-colinearity. If all the variables are correlated, it will become difficult for the model to tell how strongly a particular variable affects the target since all the variables are related. In such a case, the coefficient of a regression model will not convey the correct information.

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

   “Weathersit” has the highest correlation with target variable “cnt”

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

   We ran the model on the test set with model.predict() method to predict the data. And then used r2_score regression score function to determine the coefficient. Since the R-squared-adjust value(73.4%) of trained model and r2_score (74.7%) were very close, We conclude that our assumption of Linear Regression was correct.

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

General Subjective Questions

1. Explain the linear regression algorithm in detail.

    1. Reading and understanding the data

        Importing required libraries like pandas & numpy for data analysis and manipulation and seaborn & matplotlib for data visualization

        Cleaning and manipulating data to make it up to the standards that exploratory data analysis can be performed by treating null values if any, updating to necessary formats, changing data types if needed, removing unwanted rows or columns etc.

    2. Visualizing the data (Exploratory Data Analysis)

        Visualizing numerical variables using scatter or pairplots in order to interpret business /domain inferences.

        Visualizing categorical variables using barplots or boxplots in order to interpret business/domain inferences

    3. Data Preparation

        Converting categorical variables with varying degrees of levels into dummy variables (numerical in nature) so that these variables can be represented during model building in order to contribute to the best fitted line for the purpose of better prediction

    4. Splitting the data into training and test sets

Splitting the data into two sections in order to train a subset of dataset to generate a trained (fitted) line. Generally, the train-test split ratio is 70:30 or 80:20.

Rescaling the model using MinMaxScaling or Standardization

5. Building a model

Start the Model building with combination of Recursive Feature Elimination + Manual Feature Elimination.

First Notice the significance of P-value and VIF of Variable

1st step. Variable with High P Value and High VIF – Drop

2nd step Variable with high P Value and low VIF – Drop

3rd step Variable with low P value and high VIF -- Drop

4th step Variable with low P Value and Low VIF – are Significant

6. Residual analysis of the train data

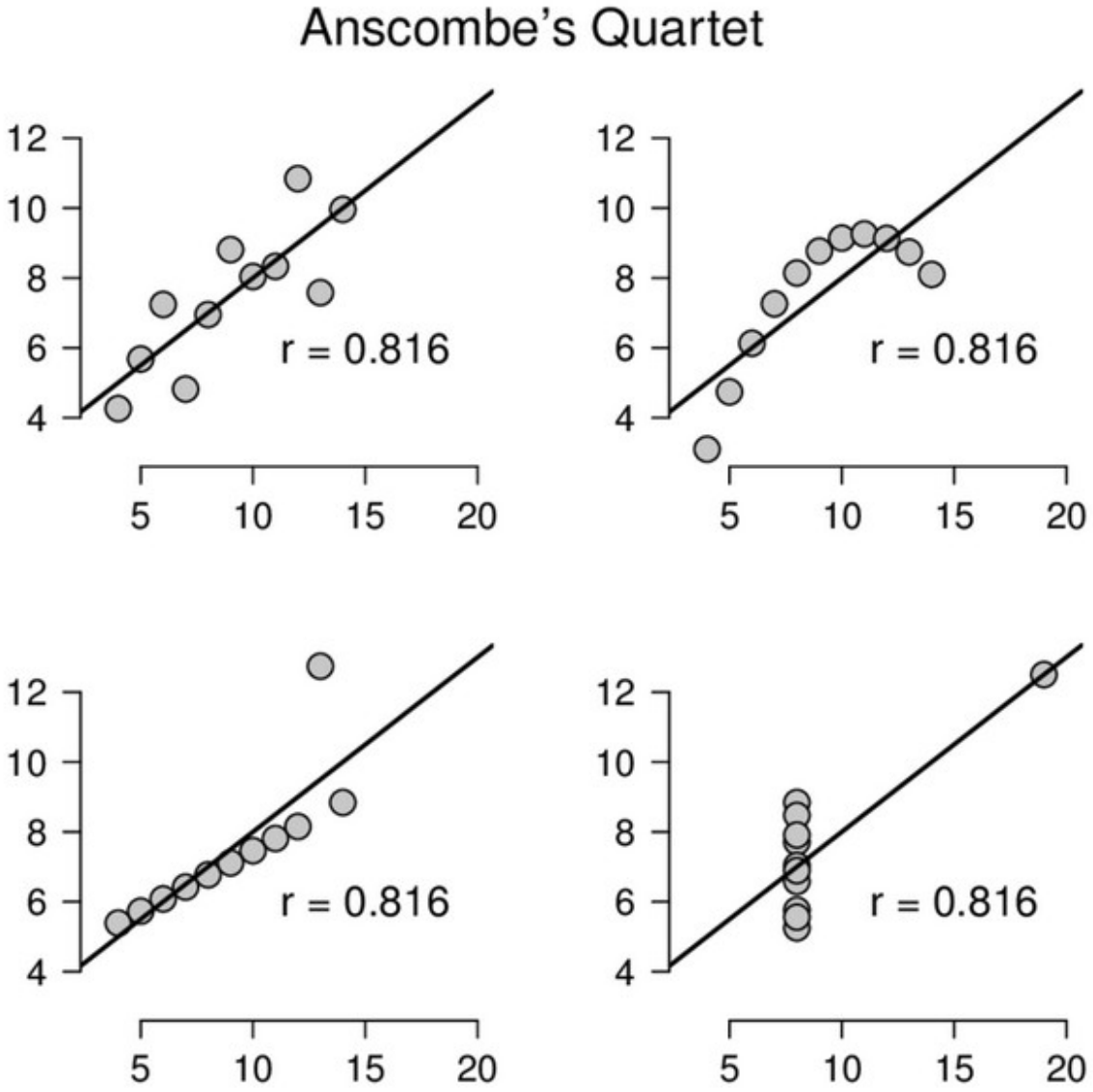
It tells us how much the errors (y_actual — y_pred) are distributed across the model. A good residual analysis will signify that the mean is centred around 0

7. Making predictions using the final model and evaluation

predict the test dataset by transforming it onto the trained dataset
Divide the test sets into X_test and y_test and calculate r2_score of test set. The train and test set should have similar r2_score
A difference of 2–3% between r2_score of train and test score is acceptable as per the standards

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet is the most appropriate demonstration which shows the danger of relying on only the Summary Statistic,
Below is shown the Anscombe's Quartet



There are four data set and each data set consists of 11 (x,y) pairs.
Each data set looks statistically identical with the correlation between x and y for each data set coming to be 0.816.
But when we plot these 4 data sets then all 4 plots looks very different.

Now we see the real relationships in the datasets start to emerge. Dataset I consists of a set of points that appear to follow a rough linear relationship with some variance. Dataset II fits a neat curve but

doesn’t follow a linear relationship. Dataset III looks like a tight linear relationship between *x* and *y*, except for one large outlier. Dataset IV looks like *x* remains constant, except for one outlier as well.

Hence it’s very important to visualize the data to get a clear picture of the distribution.

3. What is Pearson’s R?
   Pearson’s correlation coefficient R is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

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

   Feature scaling is a method used to normalize the range of independent variables or features of data.

   Since Linear regression algorithm uses gradient descent as an optimisation Technique, the required data has to be scaled.

   When we have a multiple independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons
   1. Ease of interpretation
   2. Faster convergence for gradient descent methods.

   Difference between Normalization and Standardization

Sr.	Normalization	Standardization
1.	Min & Max values used for scaling	Mean and standard deviation used for scaling

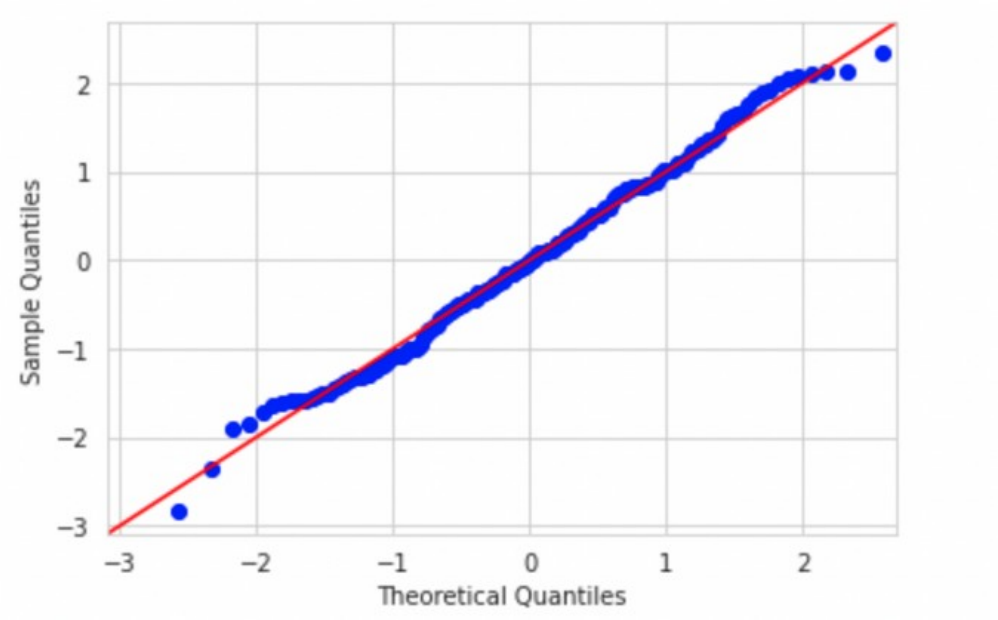
2.	It is used when feature are of different scale	It is used when we want to ensure zero mean and unit standard deviation
3.	Scale is bound between 0,1 or -1,1	Scale is not bound to any range
4.	More effected by outliers	Less effected by outliers
5.	Here MinMaxScaler is used from Sklearn	StandardScaler is used from Sklearn

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

   VIF is infinity when there is perfect correlation. This is when there is perfect correlation between two independent variables. Hence we get R2 =1, which lead to 1/(1-R2) infinity.

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

   Q-Q Plots is Quantile-Quantile 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.



Q-Q plots is very useful to determine

1. If two populations are of the same distribution
2. If residuals follow a normal distribution.
3. Skewness of distribution

Importance of Q-Q Plot

By checking distribution of the error terms or prediction error using a Q-Q plot. We can observe if any significant deviation from the mean, by checking the distribution of feature variable and consider transforming them into a normal shape.