Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 1 goes below this line> (Do not edit)

By looking at the data dictionary, I have taken 'holiday', 'workingday', 'weekday', 'weather', 'month', 'season' and 'year' (All renamed) as categorical, and their box plots infer that :

1. Bike demand shows strong seasonality with the **months** May to October or Summer and Fall **season**s having higher Bike counts. Highest for Jun-sep and fall season.
2. **weather** being good (clear with less clouds) is a driver for higher bike counts
3. The day being a **holiday** impacts the bike demand negatively (looking at the mean, not the spread or stdev)
4. **year** 2019 has higher bike demand than 2018. This was already mentioned in problem statement and also inferred from the box plot

---

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)

Thumb rule for dummy variables is : No of categories – 1

As an example, the function to create dummy variables for season will create 4 dummy variables but only 3 are enough to indicate which season it is and the fourth can be inferred, effectively used as a reference category. Using the 4th can create multicollinearity issues due to the above stated reason.

Hence drop_first – True should be used so as to drop the extra dummy variable we do not need.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)
The variable **temperature (temp)** seems to have the highest correlation with bike_count (cnt). Note that Prior to plotting, I had eliminated **atemp** because atemp was highly correlated with temp. Had I retained this variable, this would also have a high correlation with cnt.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

Following assumptions have been checked :

Linearity and Homoscedasticity – See plot of residuals and predicted values where the residuals are randomly scattered around 0 which conveys that the relationship between the independent and dependent variables is linear; Also Heteroscedasticity is not an issue because the values are randomly dispersed and there is no predominant funnel shape.
Independence - If we take Autocorrelation as a proxy for Independence, the Durbin-Watson value of 1.95 is reasonable close to 2 indicating no autocolinearity.
Normality – See plot of Residual (Error term) histogram where residuals are normally distributed with mean around 0 validating this assumption
Multicollinearity – Ignoring the constant, all the VIF values for X_train_final are well below 5 which says that there is no multicolinearity

---

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

My final model equation is :

const*0.41 + year*0.23 + temperature*0.53 + humidity*-0.33 + windspeed*-0.22 + month_jul*-0.09 + month_sep*0.05 + season_spring*-0.1 + season_winter*0.06

According to me, **temperature**, **humidity** and **windspeed** contribute significantly towards explaining the demand. Note that although **year** variable has higher coefficient than windspeed, I am not mentioning this because the same is already explained in the problem statement, "bike-sharing systems are slowly gaining popularity, the demand for these bikes is increasing every year proving that the column 'yr' might be a good variable for prediction". I have mentioned top 3 which are inferred after model building. Higher temperatures coupled with lower humidity and windspeed contributes to higher demand for bikes in general.

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Linear regression is a fundamental statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (predictors). It is the most simple Supervised Machine learning algorithm. It is of 2 types. Simple LR and Multiple LR.

Simple Linear Regression follows **y = mx + c** model where y is to be predicted (what we are modelling for) and x is the independent variable, m is the coefficient and c is the constant. This is a straight line when plotted.

Multiple Linear regression simply is an extension of Simple LR where multiple predictors, each with different coefficients are modelled . The equation **y = mx + ny + oz +c** for example says that x,y and z in some proportion and magnitude determine y depending on the magnitude of coefficients. Usually real world data is Multiple LR problem.

Following are the steps for Linear regression :

Data Preprocessing:

Handle missing values.

Encode categorical variables.

Scale features if necessary.

Model Building:

Fit the Model- Use the least squares method to estimate the coefficients Beta

The goal is to minimize the sum of the squared differences between the observed values and the values predicted by the model. The cost function (Mean Squared Error, MSE) is minimized.

Model Evaluation:

R-squared: Measures the proportion of the variance in the dependent variable that is predictable from the independent variables.

Adjusted R-squared: Adjusts the R-squared value based on the number of predictors.

Residual Analysis: Check the residuals to ensure they are normally distributed and homoscedastic (constant variance).

Model Interpretation:

Coefficients: Interpret the coefficients to understand the relationship between each predictor and the target variable. P-values: Assess the statistical significance of each predictor.

Prediction:

Use the fitted model to make predictions on new data.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed. It illustrates the importance of graphing data before analyzing it and to show how statistical properties can be misleading if not visualized.

Key Points of Anscombe's Quartet:

Identical Statistical Properties:

1. Each dataset has the same mean for both (x) and (y).
2. Each dataset has the same variance for both (x) and (y).
3. Each dataset has the same correlation coefficient between (x) and (y) (approximately 0.816).
4. Each dataset has the same linear regression line: (y = 3 + 0.5x).

Different Graphical Representations:

Despite having identical statistical properties, the datasets look very different when plotted, highlighting the importance of visualizing data.

The Four Datasets:

Dataset 1: A typical linear relationship.

Dataset 2: A perfect quadratic relationship.

Dataset 3: A linear relationship with an outlier.

Dataset 4: A vertical line with an outlier.

Importance of Anscombe's Quartet:

Illustrates the Limitations of Summary Statistics: It shows that datasets with identical summary statistics can have very different distributions and relationships.

Emphasizes Data Visualization: Highlights the importance of visualizing data to uncover patterns, relationships, and anomalies that summary statistics might miss.

Question 8. What is Pearson's R? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

Pearson's R is a measure of the linear relationship between two variables. It quantifies the degree to which two variables are linearly related, providing both the strength and direction of the relationship.

$$\boldsymbol{r} = \frac{\sum(\boldsymbol{X_i} - \bar{\boldsymbol{X}})(\boldsymbol{Y_i} - \bar{\boldsymbol{Y}})}{\sqrt{\sum(\boldsymbol{X_i} - \bar{\boldsymbol{X}})^2 \sum(\boldsymbol{Y_i} - \bar{\boldsymbol{Y}})^2}}$$

Pearson's R in linear regression is often used to assess the strength and direction of the linear relationship between the independent and dependent variables before building a regression model. It can also help identify multicollinearity by checking the correlation between independent variables

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

1: Perfect positive linear relationship.

-1: Perfect negative linear relationship.

0: No linear relationship.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

Scaling is the process of transforming the features of your data so that they are on a similar scale. Scaling Improves model Performance, prevents dominance and enhances interpretability. Linear regression models perform better when the features are on a similar scale. This can lead to faster convergence and better model performance. Another reason to prefer scaling is that if not done, features with larger scales can dominate the learning process, leading to biased results. Scaling ensures that each feature contributes equally to the model. Scaled data also can make the results of the model more interpretable, especially when comparing coefficients in linear models.

I have used Min-Max scaling in the code as this is simpler to use and sensitive to outliers and transforms the data to a fixed range, in my case, 0 to 1. This method is used when we know the data distribution and want to preserve the relationships between data points.

There is also Standardised scaling where the data transforms into a scaled data a mean of 0 and a standard deviation of 1. This is less sensitive to oputliers and useful when the data follows a Gaussian distribution or when you want to standardize the data for algorithms that assume normally distributed data.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

As observed from my code, X_train_rfe1 (the train data retained after RFE stage) has 4 variables whose VIF came to infinity. I attribute this to some variables or features being redundant or completely interdependent (complete multicollinearity). One or more of these variables can be perfectly predicted by a linear combination of the other variables. Example in my case holiday, workingday, weekday_sat, weekday_sun can probably be explained by other variables which have not so high VIF.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

A Q-Q plot is used to check whether a dataset follows the normal distribution. In the context of linear regression, it is used to check the normality of residuals. One of the key assumptions of linear regression is that the residuals (errors) are normally distributed. Here’s why this is important:

Normality of Residuals:
Assumption: The residuals should be normally distributed for the results of hypothesis tests (e.g., t-tests for coefficients) to be valid.
Check: A Q-Q plot helps visually assess whether the residuals deviate from normality.
Identifying Deviations:
Heavy Tails: If the points on the Q-Q plot form an S-shaped curve, it indicates heavy tails (more extreme values than expected).
Skewness: If the points deviate from the line in a systematic way, it indicates skewness (asymmetry in the distribution).
A QQ plot helps in model validation by ensuring that the assumption of normality of residuals is

met, which is important for the validity of statistical tests and confidence intervals. I have not used any QQ plots in my analysis as Histogram of residuals was enough in my case.