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)

1. In clear weather, demand is higher. in Mist or cloudy weather, demand is moderate and it is very less in case of rain/snow or thunderstorm.
2. Sequence of demand as per season is: fall > summer > winter > spring
3. Year 2019 has higher demands as compared to 2018.

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)

While creating dummy variables, let's say for season (summer, spring, winter, fall), 4 dummy variables are created if we don't use **drop_first=True**. While in reality, 3 dummy variables are sufficient to have impact of all 4 values. For example, summer – 100, spring – 010, winter – 001, fall – 000. We don't need to add a new feature for fall, 3 new features (summer, spring, winter) are sufficient. This increases model performance and reduces complexity.

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)

'temp' and 'atemp' are highly correlated with the target variable = 0.63 both.

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)

By residual analysis, made a histogram of the errors and it followed normal distribution with mean error 0 which satisfies the assumption of linear regression.

Created a graph which shows variance of errors among all the predicted y values and it gives almost constant variance (homoscedasticity) and no pattern followed.

Removed all the multicollinearity from the model.

Hence the model now follows all its required assumptions.

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)

'yr', 'mnth', 'season_spring'

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 supervised leaning algorithm used to predict outputs for continuous output variable and to get insights which features are affecting the output variable and how much those features affect the target variable.

Assumptions of Linear Regression:
Some assumptions need to be considered while making Linear regression model:

1. Input variable and target variable should have a linear relationship.
2. Input features should be independent of each other.
3. Errors should follow normal distribution with mean error 0
4. Errors should have constant variance (Homoscedasticity)
5. Errors should be independent to each other.

There are two methods based on which the best fit line or the best coefficients for the input features are decided for linear equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots\cdots + \beta_n x_n$ in which errors are minimized:

1. **OLS (Ordinary Least Square)** – sum of squared errors are minimized which is sum of square of the difference between actual y and predicted y.
2. **Gradient Descent** – This is an iterative method in which cost function is minimized by converging the derivative of the cost function to 0 using a step size also called learning rate.

Metrics:

1. **R^2**: It tells how much y is explained by X, range is 0 to 1. $R^2 = 0$ tells there is no relationship, $R^2 = 0.5$ tells moderate relationship and as it goes towards 1, which tells a strong relationship.
2. **Adjusted R^2**: In case of multiple linear regression, when features are added to make the model, R^2 always increases irrelevant of the significance of the added feature or it remains same when coefficient is 0. But Adjusted R^2 is penalized by no of features. So, if irrelevant feature is added, Adjusted R^2 may decrease. Hence Adjusted R^2 is a better measure in case of MLR.
3. **VIF (Variance Inflation Factor)**: This is used to check multicollinearity between the input features excluding output variable. VIF>10, generally considered high which means the particular column is highly correlated to some of other columns. VIF>5, need to inspect that column, VIF<5 is considered good.
4. **P-value**: if p-value of a feature is <0.05, feature is considered significant. This is based on a hypothesis testing with null hypothesis $\beta_0 = 0$. Hence if p-value < 0.05, we can reject the null hypothesis and can say $\beta_0 \neq 0$ and the feature is significant.
5. **F-statistic or Prob(F-statistic)**: This is a metric which tells about the overall model fit. If F-statistic is higher and Prob(F-statistic) < 0.05, we can say that overall model is a good fit.
6. **RMSE (Root Mean Squared Error)**: This is square root of average of the sum of squared errors (square root of average squared difference between actual and predicted values). This is used to compare the models which model is better. The model with lesser RMSE is considered better.

Model Making:

1. **Dummy variables**: Create dummy variables for categorical columns as linear regression doesn't understand categories. If a column has k unique values, dummy variables will be k-1.

2. **Train Test Split**: Split the data in train and test set in 7:3 or 8:2 or as per the requirement.
3. **Feature Scaling**: Features are generally required to be scaled to have same range across all the features in linear regression. We can use either Min-Max-Scaling which converts all the values of a feature between 0 & 1 or we can use Standard-Scaling which coverts values as mean=0 and sigma=1. Feature scaling should be done after train test split.
4. We fit_transform the scaler for X_train and only transform it for X_test as fit learns the required parameters that should only be learnt from training data; hence we don't use fit in case of test data.
5. Make the model, predict values, check p-values, VIF, F-statistic, R^2, Adjusted R^2.
6. Compare R^2, RMSE of train data and test data. If there is not much difference then it is a balanced model. If test data has much higher errors than train set, then the model can be overfit or underfit.
7. **Overfitting**: When model is trained more or model is complex which is giving higher accuracy in train set and lower accuracy in test set, that means model has learned all the variances and noise from train data which does not fit to the test data well, that model is considered overfit.
8. **Underfitting**: When model is too simple and accuracy is low in case of train set as well, model is considered underfit.
9. **Residual Analysis**: Validate assumptions of linear regression related to errors.

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)

It contains four relationship graphs between x & y.

1. First contains datapoints scattered around a linear line which is considered a good fit as line captures all the data points well. This means x & y datapoints are in a linear relationship.
2. In second (Top right corner), datapoints are scattered as polynomial for which straight line is not a good fit. This means x & y datapoints are not in a linear relationship.
3. In third, datapoints are in one line but one point is an outlier and located at very different position, due to which the regression line has been changed. So, outliers affect the model, we should remove the outliers before making model.
4. In forth, at one x, there are multiple y values which tells there is no relationship between x & y till now. But there is one point which is far away and at different x, a regression line has been fit now which in reality does not make any sense. So outliers affect the model badly.

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 correlation coefficient between two variables. It tells the direction as well as strength of the relationship. Range is -1 to +1. -1 tells perfect negative correlation, 0 tells no correlation, +1 tells high positive correlation.
R = Cov(x, y) / sigma(x) . sigma(y)

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)

Features in a dataframe are generally required to be scaled to have same range across all the features to fit in a linear regression model so that interpretation of coefficients become easier. We can use either **Min-Max-Scaling** which converts all the values of a feature between 0 & 1 (Formaula is (x – min(x)) / (max(x) – min(x)) or we can use **Standard-Scaling** which coverts values as mean=0 and sigma=1 (Formal is (x - μ) / sigma). Feature Min-Max-Scaling is called **Normalization** and Standard-Scaling is **Standardisation**. Scaling should be done after train test split.

We fit_transform the sacaler for X_train and only transform it for X_test as fit learns the required parameters that should only be learnt from training data; hence we don't use fit in case of test 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)

When there is perfect multicollinearity between one variable and one or more other variables, R^2 becomes 1, and thus VIF becomes infinite. We can interpret this with the following formula:

VIF = 1 / (1 - R^2)

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)

Q-Q plot is used to check if the errors are normally distributed or not which is one of the assumptions of linear regression. It is a quantile-quantile plot where at X-axis contains theoretical quantiles from the normal distribution and Y-axis contains quantiles from the dataset. And if the datapoints are closely aligned with the 45° line, that is considered as errors following normal distribution.
