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)

According to the analysis, following are the Categorical variables: 'season', 'yr', 'mnth', 'holiday', 'weekday','workingday', 'weathersit'.

From their univariate analysis, There are four seasons, 2 years in considerations, 12 monts in consideration and 4 weathers in considerations.

From the bivariate analysis between categorical variables and count, following observarions are made,

1. Summer, Fall and winter seasons have more bike sharing, highest in fall
2. In 2019 the bike sharing is increased
3. From May to September month overall bike sharing increases
4. Holiday median is lower than non-holiday, i.e. bike sharing seems to be lower on holidays
5. Nothing specific on the basis of week days, all the weekdays seems to have similar tendency.
6. Bike sharing is more in the clear weather, then in Mysty weather and they avoid in case of rains

---

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)

In case of regression analysis, categorical need to be converted in dummy variables according to their category captions. It basically does the Boolean conversions for the category values. But in order to reduce the redundancy categorical variables are created with n-1 rule where n represents the number of values.

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 have more correlation with the target variable cnt according to the pairplot and heatmap.

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)

1) There should be linear relationship between dependent and independent variables:
   Plot regression plot between the dependent and independent variables. Y predicted

depicts the equation created using dependent variables. The plot is linear. And residuals are randomly scattered around the line.

2) Independent of errors: Durbin Watson test suggests if the plot is independent of the errors or not. Value around 2 suggest that there is no autocorrelation. From the model the Durbin Watson is having value of 2.094. which is around 2 and means that there is no evidence of autocorrelation.
3) The error should be normally distributed:
from the plot between predicted values and actual values (Residuals), conclusion is that the error is normally distributed around 0.
4) Homoscedasticity: The variance of the error should be constant across all columns: Minimum standard error is ranging from 0.008 to 0.037. The error distribution is normal across the values of dependent variables.
5) No multicollinearity: VIF is the standard for multicollinearity. Maximum VIF for predictors is 2.07 which is much below 5. Inferring no multicollinearity between predictors.

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)

The impact of the feature is more if the coefficient of the feature is higher. According to that below three are the top three features. Light rain impacts negatively whereas temperature and year impact the bike sharing positively.

weathersit_Light_Rain	-0.242951
yr	0.229130
temp	0.567148

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)

<Your answer for Question 6 goes here>

Linear regression is the process for building the model of the continuous/numeric variables. The basic principle is to find the linear relationship between the dependent and independent variables.

1) Equation: Y = B0+B1*X where B0 is constant and B1 is coefficient of X.
2) The primary method to build the model of linear regression is using **Least square method**. The line where residual distance between line and all the datapoints is least that line is considered as the best fit.
3) The fitness of the model is calculated using R-Squared method.

R-Squared = (1- (RSS/TSS)) where RSS is residual sum of squares and TSS is total sum of squares.
TSS is the average line of average is independent of the X variables. R-Squared is the major of the fitness for all the variables.
Along with R-Squared value the P value or the significance value also important to check as it will help to understand if the feature is significant or not.

4) F-Statistics: This is the parameter used for checking fitness of the model. Instead of checking the values of individual coefficients the evaluation is done on the total model.

Assumptions of the linear regression:

1) Linear relationship between dependent and independent variables
2) Error terms are normally distributed
3) Error terms are independent of each other
4) Error terms have constant variance

In case of multilinear model the equation becomes,
Y= B0+B1*X1+B2*X2……

Where X1,X2 etc. are also known as predictors and features.
In MLR model the feature selection becomes very important. The model should not be overfit.
There should not multicollinearity between features.
For the selection of the features RFE (Recursive feature elimination) can be used which is an automated process for identifying the top number of variables for the model.
Manual method also can be used which is basically checking and adding the columns one by one in the model.
For the multicollinearity check we use VIF (variance inflation factor) = 1/(1-R^2).
If the VIF value is more than 10 then it should be rejected.
If it is more than 5 then it should be tested and if the value is less than 5 then the feature is not having multicollinearity.

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)

<Your answer for Question 7 goes here>
Anscombe quartet consists of four datasets for which the descriptive statistics are identical but their visual representations and distributions are different. This was demonstrated by statistician Anscombe.
Key characteristics:

1) Same descriptive statistics
2) Distinct visualisations

The significance of this quartet underlines the importance of Exploratory data analysis, data visualizations and limitations of statistical summaries.

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)

<Your answer for Question 8 goes here>
Pearson's R is a valuable statistical measure for assessing the strength and direction of linear relationships between two continuous variables.

The range is between -1 to 1.
Formula:
$r = (\Sigma[(x_i - \bar{x})(y_i - \bar{y})]) / (\text{square-root}(\Sigma(x_i - \bar{x})^2) * \text{square-root}(\Sigma(y_i - \bar{y})^2))$
where,
x_i – individual datapoint of first variable
$\bar{x}$ – is the mean of the datapoints of the first variable
y_i – individual datapoint of second variable
$\bar{y}$ – is the mean of the datapoints of the second variable

It is being used in the linear regression model mainly with the features having high correlation coefficients
Cons:
1) It assumes that the relation is linear
2) It is very sensitive to outliers.

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)

<Your answer for Question 9 goes here>
- Scaling refers to the method of bringing all numeric features on a same scale between 0 and .
- If different variables are of different ranges then their coefficients will vary accordingly which will lead to misinterpretation of the algorithm as well as equation. Bringing them on same scale makes the algorithm understandable.
- There are mainly two scaling methodologies: min-max scaling also known as normalized scaling and standardized scaling.
- Formulae for min-max scaling: X_scaled = (X - X_min) / (X_max - X_min)
- Formulae for standardized scaling: X_scaled = (X - mean) / standard deviation

The difference between both is that min-max brings features in the range of 0 and 1, whereas standardized brings everything at the mean 0 and standard deviation 1.
Standardized scaling is less affected by outliers than normalized scaling.

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)

<Your answer for Question 10 goes here>
VIF is variance inflation factor and calculated as,
VIF = (1/(1-R^2)
VIF is infinite means denominator of the above equation is 0. That means R-square value is equal to 1
R-Squared = 1-(RSS/TSS) where RSS – Residual sum of squares and TSS means Total sum of squares. R-Squared can be 1 **only if RSS is 0. i.e. all the datapoints are on the line.** This can happen

if model is overfit or there is multicollinearity between features.

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)

<Your answer for Question 11 goes here>

Q-Q plot is the plot in the linear regression, the sample data is compared with the theoretical datapoints. Quantile to quantile data comparison is made.

1) The data for both the datasets is first sorted
2) Quantiles are plotted against each other
3) If there is deviation of plotted line from the theoretical line then the data might not follow the theoretical line.

It is mainly used for,

1) the error normality assumption checks.
2) Outliers checks
3) Heavy tails

In python code:

```
import pylab as py
sm.qqplot(res_np,line = 's')
```
