Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer abouttheir effect on the dependent variable?** (3 marks)

Ans. 1 – Following are the observations of the effect of categorical variables on the dependent variable:

- Total rental bikes seem to be increasing in Summer and Fall (may be due to pleasant weather during these seasons in US)
- Total rental bikes are on the rise from 2018 to 2019
- Total rental bikes generally increases from January till around September and then declines till December
- Total rental bikes has a lower median during holidays. In some holidays people use the bike sharing more and in some holidays quite less (might be people are using it to commute during workdays more)
- No specific pattern by weekdays
- Slightly higher median on a working day (but nothing significant)
- The bike rentals are higher when the weather is - "Clear, Few clouds, Partly cloudy, Partly cloudy" and lowest when "Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds"
- As the precipitation increases, the bike rentals decreases

2. **Why is it important to use “drop_first=True” during dummy variable creation?** (2 mark)

Ans. 2 – Using “drop_first=True” is important during dummy variable creation because it helps in reducing the extra columns created during the dummy variable creation.

Also, it reduces the correlation between the dummy variables since one of the categories is omitted during the dummy variable creation.

Example : if there is a categorical variable named as “product_type” and it contains 3 categories – ‘electronics’, ‘food’ and ‘clothes’, then when we drop_first=True while creating dummy variables, the first category is taken as the base category and is omitted while making the dummy variables.

So, we would have 2 dummy variables created out of the product_type column.

- Product_type_food – set as 1 where ever the category is food else 0
- Product_type_clothes – set as 1 where ever the category is clothes else 0
- And where ever both of the above columns are 0, that represents category ‘electronics’

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?** (1 mark)

Ans. 3 – Both '**temp'** and '**atemp'** are highly correlated with the target variable 'cnt'.

Both have the correlation coefficient of 0.63 with 'cnt'.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?** (3 marks)

Ans. 4 – After the final model was confirmed, Residual analysis was performed on the train data to validate the assumptions of Linear Regression:

Residual analysis steps:

- Predictor variables from the train data are passed into the final model to predict the respective target variable
- Error terms (residuals) are calculated (Actual target value – predicted target value)
- Histogram is plotted for the calculated error terms
- Scatter plot is plotted for the error terms

Assumptions of linear regression:

- Error terms are normally distributed and mean = 0 :
  - Using the histogram of the error terms, it was observed that the histogram followed Normal distribution and was centred around 0 (i.e. mean = 0).
  - Hence the assumption is validated
- Error terms are independent of each other :
  - In the scatter plot of error terms, No specific pattern was observed. Thus proving that the error terms are independent.
  - Hence the assumption is validated
- Error terms follow the same variance for all the values of predictor X (Homoscedasticity):
  - In the scatter plot of the error terms, the variance (spread) of the error terms did not vary significantly as the X value increased. Thus same variance was observed at any value of X.
  - Hence the assumption is validated

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

Ans. 5 – Based on the final model, the top 2 features for explaining the demand of shared bikes are:

- temperature
- is Weather : "Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds"
- year

These are most significant based on their coefficients in the linear model. These variables have the highest coefficients.

Since, the coefficients also represent weightages, a higher coefficient means higher weightage or higher change in the Target variable with 1 unit change in the respective X variable.

General Subjective Questions

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

Ans. 1 - Linear regression is a statistical and machine learning algorithm used to model the relationship between a dependent (target) variable and one or more independent (input) variables by fitting a linear equation to the observed data.

It is used when we need to predict a continuous variable.

- Objective of Linear Regression:

  The objective of the linear regression is to predict the value of the target variable assuming linear relationship between the Predictor variables and the target variable.

  The main aim is to come up with a linear equation which best describes the observed data.

  y = m1 X1 + m2 X2 + ………. + mn Xn + C

  y - target variable

  m - coefficients

  X - predictor variable

  C - intercept

  To get the best fit line, the algorithm needs to get the optimum coefficients for the X variables.

- Linear Regression Algorithm:

  The algorithm works by finding the best fit line through the observed data.

  This is done by getting the most optimum values of the coefficients of the X variables.

  And in turn, this is done by minimizing the Cost function, which in linear regression's case is the Mean squared error (MSE)

  MSE – mean of squared difference between actual values and the predicted values

  The cost function can be minimized using one of the following ways:

  - **Differentiation method** : the concept followed is that at the minima, the first order derivative of the cost function will be zero.

  - **Gradient Descent method** : this is an iterative optimization algorithm used to minimize the cost function.

- Predicting Target variable:

  After the best fit line with the optimum coefficients is calculated, predictions are made using this line.

  For a particular X value, a corresponding value lying on this line is taken as a prediction.

- Model Evaluation:

  The best fit line is evaluated based on some metrics:

  - **R-sqaured** : this measures the amount of variance in target variable explained by the predictor variables. More the R-squared the better. But need to check for Over-fitting and settle for an optimum R-squared.
  - **RMSE** (Root mean Squared error)

- Assumptions of Linear Regression:

  - Assumes a linear relationship between the target and predictor variables
  - Error terms (residuals) are normally distributed with mean = 0
  - Error terms are independent of each other
  - Error terms are having constant variance across all the values of independent variables (Homoscedasticity is followed)

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

Ans. 2 – Anscombe's quartet was created by statistician Francis Anscombe in 1973.

It comprises of four datasets with similar summary/descriptive statistics like mean, variance, correlations and linear regression lines, but completely different when represented visually.

The main purpose of Anscombe's quartet is to show the limitations of solely relying on the summary statistics numbers and to show the importance of observing the datasets visually in EDA.

Each of the four datasets in Anscombe's quartet has 11 X-Y pairs. When they are plotted, each of them seems to have a unique X-Y pattern, correlation and variation. But the summary statistics come out to be same.

Thus we should always visualize dataset for better understanding.

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

Ans. 3 - Pearson Correlation Coefficient(PCC), denoted as r, measures the strength and the direction of the linear relationship between two continuous variables.

It ranges from -1 to 1. PCC > 0 would show a positive relationship, meaning one variable increases with increase in another. PCC < 0 would show a negative relationship, meaning one variable decreases with the increase in another.

Example:

- Speed and time are negatively corelated, since if you increase speed, the time it takes to cover a distance decreases.
- Height and weight are positively corelated, since weight increases with the height of a person.

r = 1 : Perfect positive linear relationship
r = 0 : No linear relationship
r = -1 : Perfect negative linear relationship

For PCC, the variables should follow a normal distribution and should be homoscedasatic.
Also, extreme outliers affect the PCC.
Pearson's r is mostly used in linear regression to find out the degree of linear relationship between the target variable and the predictor variables, and also among the predictor variables.

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

Ans.4 – Scaling:

Scaling in linear regression is a data preparation step where the independent variables are transformed to fit into a particular range or magnitude.

Why scaling?

Most of times the independent variables are varying in magnitude. The scaling is required for the following reasons:

- **Interpretability** : If the scaling (bringing them in the same range of magnitude) is not done, then interpreting the coefficients(weights) would be difficult. We would not be able to compare the significance of different independent variables. So scaling is required for ease of interpretation.
- **Faster convergence for gradient descent method** : If the independent variables are in the same range, it becomes easier for the gradient descent to come to the optimum line quicker.
- For some ML algorithms which are based on Euclidean distance, the different magnitudes of the variables might skew/bias the results to the higher magnitude variable.

Important note - It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalized vs Standardized scaling:

Normalization brings the data into the range of 0 and 1. Whereas Standardization converts the data into a standard normal distribution (roughly between -3 to 3), that is, mean =0 and standard deviation = 1.

Also, normalization tends to lose some info in the data since it does not capture the degree of variation in the data.

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

Ans. 5 – VIF checks the dependency of a predictor variable on the combination of all the other predictor variables.

When VIF reaches infinity, it indicates perfect multi-collinearity. This means that a predictor variable is a perfect linear combination of one or more other predictor variables.

Mathematically – VIF formula is : **VIF = 1/(1-R-squared)**
When a predictor variable can be perfectly predicted by a combination of other predictor variables, the R-squared for such relationship becomes very close to 1 (100%).
Thus, the denominator – (1-R-squared) in the above equation becomes 0 and VIF approaches infinity.

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

Ans. 6 – A Q-Q plot (Quantile-Quantile plot) is a graphical representation which helps to determine whether a set of data follows a particular distribution or not, mostly Normal distribution.

It can also be used to check is two samples came from the same population or not. In that case, a q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set(or theoretical quantiles)

In linear regression, Q-Q plot can be used to check the assumption of normality of the residuals(error terms) and also to determine any outliers in the data.

Importance of Q-Q plot:

- Helps in **checking normality of residuals**, which is necessary to carry out many statistical tests and forming confidence intervals for the regression coefficients.
- Helps in **checking model fit** – if the residuals are not normally distributed, it may indicate that the model is not capturing all the patterns in the data and a transformation of the variables might be required.
- Helps in **checking outliers** – if there are point in Q-Q plot which are far away from the reference line, these may be outliers and investigation is necessary.

The Q-Q plot in linear regression is a diagnostic tool to evaluate the normality assumption of residuals. If the plot shows substantial deviations from normality, it may indicate problems with the model, such as the presence of outliers, non-linearity, or heteroscedasticity, and guide further steps to improve model fit.