Assignment-based Subjective Questions-Answers

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

Answer : As per the below table taken after final regression model, many of the predictor variables are categorical in nature and some of them are encoded to dummy variable like month, season, weathersit .So , we can conclude that these categorical variables are statistically significant and explain the variance in model very well.

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2335	0.025	9.324	0.000	0.184	0.283
yr	0.2357	0.008	28.162	0.000	0.219	0.252
workingday	0.0219	0.009	2.455	0.014	0.004	0.039
temp	0.4178	0.029	14.354	0.000	0.361	0.475
windspeed	-0.1384	0.025	-5.438	0.000	-0.188	-0.088
spring	-0.1141	0.016	-7.189	0.000	-0.145	-0.083
winter	0.0604	0.013	4.617	0.000	0.035	0.086
weathersit_2	-0.0802	0.009	-8.984	0.000	-0.098	-0.063
weathersit_3	-0.2885	0.025	-11.477	0.000	-0.338	-0.239
month_3	0.0461	0.015	3.098	0.002	0.017	0.075
month_5	0.0435	0.016	2.718	0.007	0.012	0.075
month_9	0.0748	0.016	4.707	0.000	0.044	0.106

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

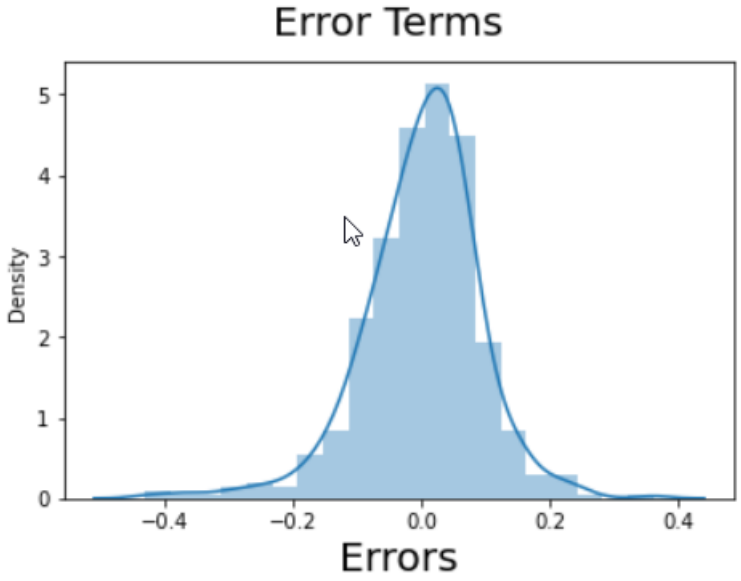
Answer: Setting drop_first=True drops the first dummy variable (column), which helps avoid the dummy variable trap by removing the redundancy.
Here's how it works:
Avoiding Redundancy: By dropping the first dummy variable, the remaining k−1 dummy variables provide the necessary information to uniquely represent each category without redundancy.
Model Interpretation: The dropped category becomes the baseline category against which the other categories are compared. This makes the model easier to interpret.

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

Answer: The pair plot shows highest correlation for registered variable having correlation 0.945. But we are not using casual and registered in our pre-processed training data for model training. casual + registered = cnt. This might leak out the crucial information and model might get overfit.
So, excluding these two variables atemp and temp is having highest correlation with target variable cnt.
As per the correlation heatmap, correlation coefficient between atemp and cnt / temp and cnt is 0.63.

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

Answer : To validate assumptions of the model, and hence the reliability for inference, we go with the following procedures:
• **Residual Analysis**: We need to check if the error terms are also normally distributed (which is in fact, one of the major assumptions of linear regression). I have plotted the histogram of the error terms and this is what it looks like:



The residuals are following the normally distribution with a mean 0.

• **Linear relationship between predictor variables and target variable:**

This is happening because all the predictor variables are statistically significant (p-values are less than 0.05). Also, R-Squared value on training set is 0.832 and adjusted R-Squared value on training set is 0.828. This means that variance in data is being explained by all these predictor variables.

• **Error terms are independent of each other**:

Handled properly in the model. The predictor variables are independent of each other. Multicollinearity issue is not there because VIF (Variance Inflation Factor) for all predictor variables are below 5.

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

Answer: Top 3 features significantly contributing towards demand of shared bikes are:

1) temp (coef: 0.4178)
2) yr (coef: 0.2357)
3) month_9 (coef: 0.0748)

General Subjective Questions-Answers

1. Explain the linear regression algorithm in detail ?

Answer: Linear regression is a fundamental statistical method used to model the relationship between a dependent variable and one or more independent variables. It is a popular technique in machine learning for predictive modeling.

Here is a detailed explanation of the linear regression algorithm:

Key Concepts:

- Dependent Variable (y): The variable we are trying to predict or explain.
- Independent Variables (X): The variables used to predict the dependent variable.
- Linear Relationship: The relationship between the dependent and independent variables is assumed to be linear.

Types of Linear Regression:

Simple Linear Regression: Models the relationship between a single independent variable and a dependent variable.

Multiple Linear Regression: Models the relationship between multiple independent variables and a dependent variable.

Objective:
The objective of linear regression is to find the best-fitting line (or hyperplane in higher dimensions) that minimizes the sum of the squared differences between the observed values and the values predicted by the model. This method is known as Ordinary Least Squares (OLS).

Steps in Linear Regression:

- **Data Preparation**:
  - Collect and clean the data.
  - Split the data into training and testing sets.
  - Normalize or standardize the data if necessary.
- **Model Initialization**:Initialize the parameters (coefficients and intercept) to some values, often zero.

- Hypothesis Function: The hypothesis function for linear regression is: y^=β0+β1X1+β2X2+…+βnXn\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_ny^=β0+β1X1+β2X2+…+βnXn where y^\hat{y}y^ is the predicted value.

- **Cost Function:** The cost function (Mean Squared Error) measures the average of the squared differences between the observed and predicted values: J(β)=12m∑i=1m(y^i−yi)2J(\beta) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}_i - y_i)^2J(β)=2m1∑i=1m(y^i−yi)2
  where mmm is the number of observations.

- **Optimization Algorithm:**The goal is to minimize the cost function. The most common method is Gradient Descent, which iteratively updates the parameters to reduce the cost function.

- **Gradient Descent Algorithm**:
  - Initialize parameters β0,β1,…,βn\beta_0, \beta_1, \ldots, \beta_nβ0,β1,…,βn.
  - Repeat until convergence:
    βj:=βj−α∂J(β)∂βj\beta_j := \beta_j - \alpha \frac{\partial J(\beta)}{\partial \beta_j}βj:=βj−α∂βj∂J(β)
    where α\alphaα is the learning rate.
  - The partial derivative of the cost function with respect to each parameter is: ∂J(β)∂βj=1m∑i=1m(y^i−yi)Xij\frac{\partial J(\beta)}{\partial \beta_j} = \frac{1}{m} \sum_{i=1}^{m} (\hat{y}_i - y_i) X_{ij}∂βj∂J(β)=m1i=1∑m(y^i−yi)Xij

  - **Model Training:**
    - Use the training data to fit the model and find the optimal parameters.

  - **Model Evaluation**: Evaluate the model using the testing data.

- Common evaluation metrics include R-squared, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

- **Prediction:** Use the trained model to make predictions on new data.

2. Explain the Anscombe's quartet in detail ?

Answer : Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet they have very different distributions and appear very different when graphed. The quartet was created by the British statistician Francis Anscombe to demonstrate the importance of graphing data before analyzing it and to illustrate the effect of outliers and the influence of linear regression.

Properties of Anscombe's Quartet : Each dataset in Anscombe's quartet has the following identical or nearly identical statistical properties:

- Mean of x values
- Mean of y values
- Variance of x values
- Variance of y values
- Correlation between x and y
- Linear regression line (slope and intercept)
- R-squared value for the linear regression

Despite these similarities, the datasets are visually distinct, illustrating that statistical properties alone do not tell the whole story of the data.

Importance of Anscombe's Quartet

Anscombe's quartet emphasizes the following key points:

1. **Graphical Analysis:** It's crucial to visualize the data through scatter plots and other graphical methods. Statistical measures alone may not reveal the true nature of the data.
2. **Outliers:** Outliers can significantly affect statistical properties and regression models. Detecting and handling outliers appropriately is essential for accurate data analysis.
3. **Data Context:** Understanding the context and distribution of the data is important for drawing valid conclusions.
4. **Statistical Literacy:** The quartet encourages a deeper understanding of statistical measures and their limitations.

3. What is Pearson's R?

Answer: Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two variables. It is a dimensionless index that ranges from -1 to 1, where:

- **1** indicates a perfect positive linear relationship,
- **0** indicates no linear relationship,
- **-1** indicates a perfect negative linear relationship.

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

Answer: Scaling is the process of transforming the features of your data so that they fall within a specific range, usually to ensure that they are comparable and to improve the performance and convergence speed of certain machine learning algorithms. It helps to avoid the dominance of one feature over others due to differences in scale.

Scaling performed because:

- **Algorithm Performance**: Many machine learning algorithms (e.g., gradient descent-based algorithms, K-nearest neighbors, support vector machines) perform better and converge faster when the data features are on a similar scale.
- **Equal Contribution**: Scaling ensures that each feature contributes equally to the result, preventing features with larger scales from dominating the learning process.
- **Normalization of Features**: Ensures that the model is not biased towards features with higher magnitudes.
- **Improving Accuracy**: Scaling can help improve the accuracy and efficiency of the model by ensuring that the algorithm treats all features equally.

Normalized Scaling vs. Standardized Scaling

Both normalized scaling and standardized scaling are techniques used to adjust the scales of features, but they do so in different ways:

Normalization scales the data to a fixed range, typically [0, 1] or [-1, 1]. This is done using the following formula:

$x' = x - \min(x) \,/\, \max(x) - \min(x)$

Purpose: It compresses all the values in the dataset to a specific range, making it useful for algorithms that rely on distances between data points, such as K-nearest neighbors or neural networks.

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

Answer: If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Answer: Q-Q Plots (Quantile-Quantile plots) are 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.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

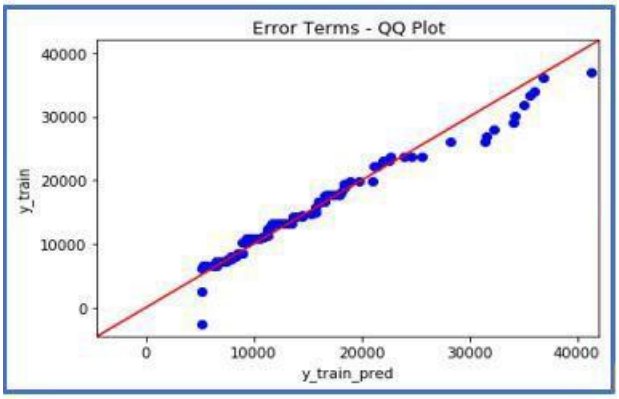
Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

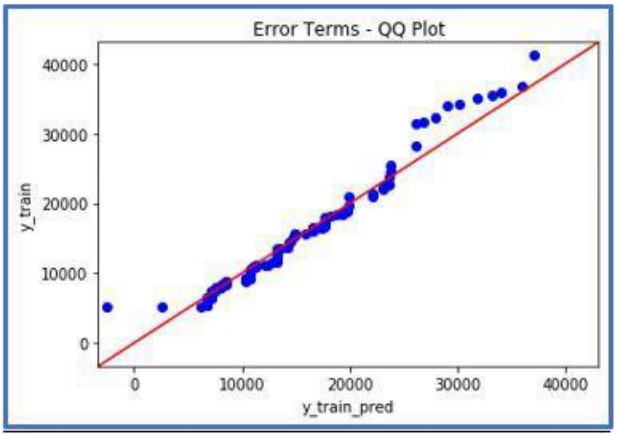
Below are the possible interpretations for two data sets.

1) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

2) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



3) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



4) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis

statsmodels.api provide qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively.