Assignment-based Subjective Question

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

Ans:

Coefficients:

```
const                        0.121392
yr                           0.233108
temp                         0.552725
windspeed                   -0.155260
season_summer                0.088176
season_winter                0.129402
mnth_sept                    0.097698
weathersit_Light_snowrain   -0.279492
weathersit_Misty            -0.078044
dtype: float64
```

Summary Interpretation:

1. **Baseline Demand:** The model's intercept is 0.12, representing the baseline demand for bikes when all other predictor variables are at their reference levels or zero.

2. **Yearly Trend:**

    - **Year ('yr'):** Each subsequent year is associated with a 0.23 unit increase in bike demand, indicating a growing trend in bike-sharing usage over time.

3. **Meteorological and Temporal Factors:**

    - **Temperature ('temp'):** A one-unit increase in temperature is linked to a 0.55 unit increase in bike demand, highlighting the importance of warmer weather in bike usage.

    - **Windspeed ('windspeed'):** An increase in wind speed is associated with a 0.16 unit decrease in bike demand, suggesting that windier conditions deter bike usage.

4. **Seasonal Impact:**

    - **Summer ('season_summer'):** The summer season sees a 0.09 unit increase in bike demand compared to the baseline season.

    - **Winter ('season_winter'):** The winter season experiences a 0.13 unit increase in bike demand compared to the baseline season, indicating a preference for bike usage in winter over the baseline season.

5. **Monthly Influence:**

    - **September ('mnth_sept'):** The month of September specifically shows a 0.10 unit increase in bike demand, signifying its popularity or favorable conditions for biking.

6. **Weather Conditions:**

- **Light Snow/Rain ('weathersit_Light_snowrain'):** Light snow or rain leads to a significant 0.28 unit decrease in bike demand, underscoring adverse weather's impact on bike-sharing.
- **Misty Conditions ('weathersit_Misty'):** Misty weather results in a 0.08 unit decrease in bike demand, indicating a mild adverse effect compared to more severe weather conditions.

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

Ans:

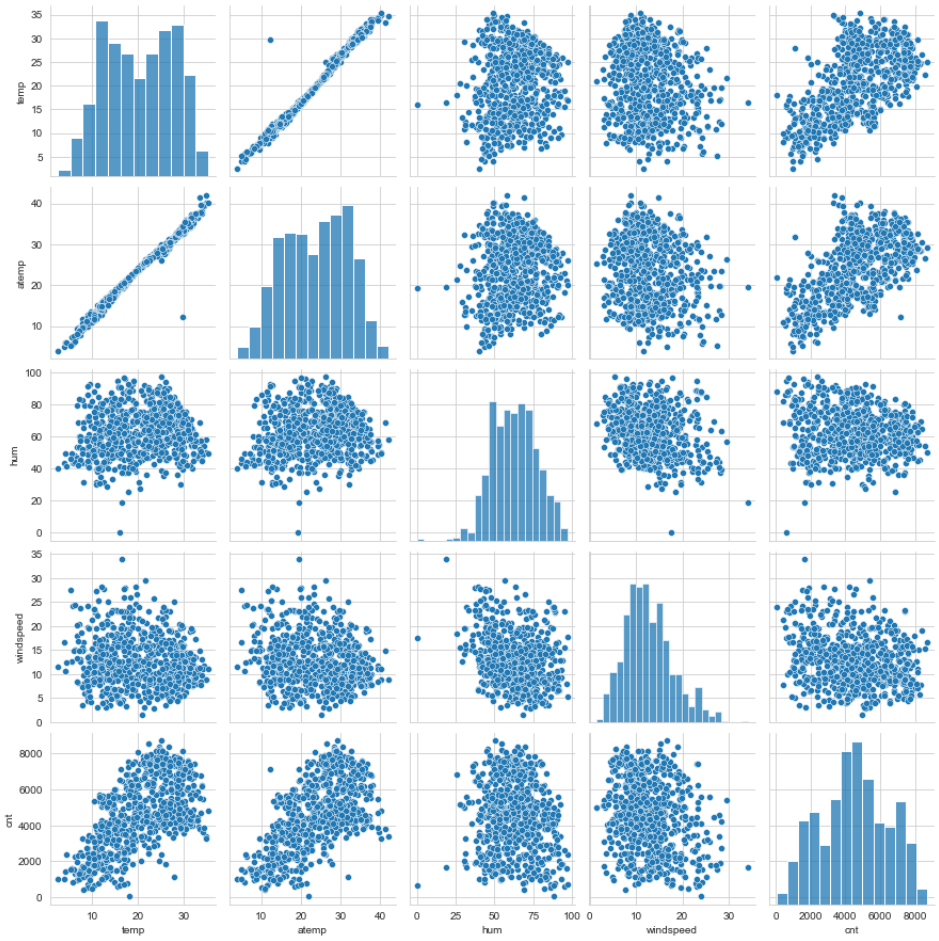
When you convert categorical variables into dummy/indicator variables (also known as one-hot encoding), you typically create a new binary (0/1) variable for each level of the categorical variable. However, this can lead to multicollinearity, which is a problem in regression models where two or more variables are highly correlated.

To avoid multicollinearity, it's important to use the drop_first=True parameter in functions like pd.get_dummies(). This parameter drops the first category level and creates dummy variables for the remaining levels. Here's why this is important:

1. **Redundancy Removal:** In the case of categorical variables, one level can always be inferred from the others.
2. **Avoiding Multicollinearity:** Including all dummy variables for a categorical variable in a regression model will result in perfect multicollinearity. This happens because the sum of all dummy variables equals one for each observation, leading to a situation where one variable can be perfectly predicted from the others. Dropping the first dummy variable helps in removing this perfect multicollinearity.
3. **Interpretation and Simplicity:** Dropping one category simplifies the model without losing interpretative power. The coefficients of the included dummy variables represent the change in the response variable relative to the dropped category. This makes it easier to interpret the coefficients in terms of differences from a baseline category.

In summary, using drop_first=True is a best practice in dummy variable creation as it helps in avoiding multicollinearity, simplifies the model, and retains the interpretability of the regression coefficients. This practice is especially crucial in linear regression models, where multicollinearity can significantly impact the estimates of the coefficients.

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



Scatter Plot Analysis:

1. The scatter plots for 'temp' and 'atemp' show a strong positive correlation with bike rentals ('cnt'), with a clear upward trend and tightly clustered points, indicating a higher demand for rentals as temperature increases.
2. The scatter plot for 'hum' (humidity) suggests a weaker, more dispersed relationship with 'cnt', implying less influence on bike rental demand.
3. The scatter plot for 'windspeed' indicates a possible negative correlation, with a more spread-out distribution of points, suggesting that higher windspeed may negatively impact rental demand.

Pearson Correlation Coefficients:

'atemp': 0.630685

'temp': 0.627044

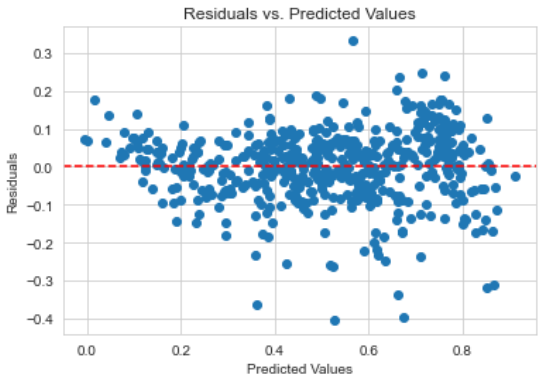
'hum': -0.098543

'windspeed': -0.235132

Summary: The pair-plot and Pearson correlation coefficients collectively indicate that 'atemp' and 'temp' are the most significant numerical predictors of bike rental demand, with 'atemp' exhibiting the highest positive correlation (approximately 0.631). This reinforces the visual analysis, which shows that warmer temperatures are favorable for bike rentals. In contrast, 'hum' and 'windspeed' have weaker negative correlations with 'cnt', suggesting that increased humidity and wind may discourage bike usage.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Linearity:



Visualizing the relationship between predicted values and residuals

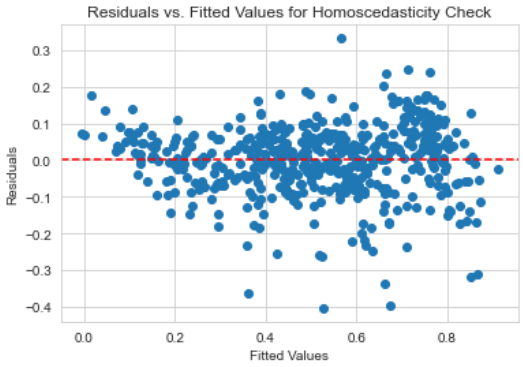
The scatter plot of residuals vs. predicted values does not exhibit any distinct patterns or systematic deviations, suggesting that the linearity assumption is reasonable for our model

Independence:

```
Durbin-Watson test: 2.0332110483834827
```

The Durbin-Watson test yields a value of approximately 2.033, which is close to the ideal value of 2, indicating a lack of autocorrelation among the residuals and supporting the independence assumption

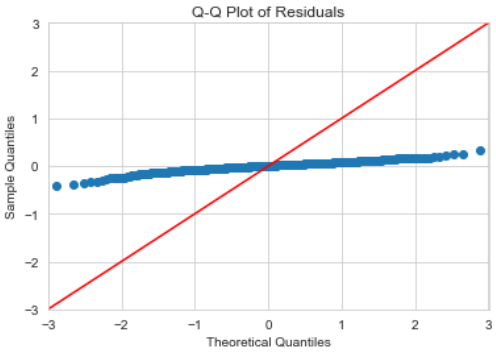
Homoscedasticity:



Visualizing the residuals vs. fitted values for constant variance

The residuals vs. fitted values plot shows a random scatter of points without a discernible pattern, which suggests that the homoscedasticity assumption holds and the residuals have constant variance across predictions

Normality of Residuals



Q-Q plot

The Q-Q plot of the residuals aligns closely with the theoretical quantiles line, with only minor deviations at the tails. This indicates that the residuals are approximately normally distributed, satisfying the normality assumption.

Multicollinearity

The VIF values for all features included in the model are well below the threshold of 5 or 10, suggesting that there is no concerning level of multicollinearity within our predictors. This is positive as it implies that our model coefficients can be estimated with a higher degree of reliability.

VIF calculation

	Feature	VIF
0	const	16.398491
1	yr	1.016143
2	temp	1.192703
3	windspeed	1.085834
4	season_summer	1.186111
5	season_winter	1.201719
6	mnth_sept	1.104313
7	weathersit_Light_snowrain	1.047105
8	weathersit_Misty	1.037445

Summary:

Our analysis indicates that the regression model fulfils the necessary assumptions for linear regression. The linearity and independence of the residuals are affirmed by visual inspection and the Durbin-Watson statistic, respectively. Homoscedasticity is supported by the lack of patterns in the residuals vs. fitted values plot, and the normal distribution of residuals is confirmed by the Q-Q plot. Lastly, multicollinearity does not appear to be a concern based on the VIF values, suggesting that our model's predictors are sufficiently independent. These validations give us confidence in the reliability of our model's findings.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.826
Model:                            OLS   Adj. R-squared:                  0.823
Method:                 Least Squares   F-statistic:                     297.4
Date:                Fri, 17 Nov 2023   Prob (F-statistic):          8.04e-185
Time:                        01:41:00   Log-Likelihood:                 484.51
No. Observations:                 510   AIC:                            -951.0
Df Residuals:                     501   BIC:                            -912.9
Df Model:                           8
Covariance Type:            nonrobust
================================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                          0.1214      0.017      7.170      0.000       0.088       0.155
yr                             0.2331      0.008     27.653      0.000       0.217       0.250
temp                           0.5527      0.020     27.313      0.000       0.513       0.592
windspeed                     -0.1553      0.026     -6.045      0.000      -0.206      -0.105
season_summer                  0.0882      0.011      8.330      0.000       0.067       0.109
season_winter                  0.1294      0.011     12.210      0.000       0.109       0.150
mnth_sept                      0.0977      0.016      6.046      0.000       0.066       0.129
weathersit_Light_snowrain     -0.2795      0.025    -11.038      0.000      -0.329      -0.230
weathersit_Misty              -0.0780      0.009     -8.701      0.000      -0.096      -0.060
==============================================================================
Omnibus:                       66.354   Durbin-Watson:                   2.033
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              143.601
Skew:                          -0.716   Prob(JB):                     6.57e-32
Kurtosis:                       5.170   Cond. No.                         10.2
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Top 3 features contributing to bike demand:
temp                        0.552725
weathersit_Light_snowrain   0.279492
yr                          0.233108
dtype: float64
```

Summary:

The regression analysis reveals that the top three features influencing the demand for shared bikes are the **temperature ('temp'), weather conditions ('weathersit_Light_snowrain'), and the year ('yr').** Temperature has the strongest positive association with bike demand, with a coefficient of 0.553, indicating that warmer temperatures significantly boost rental numbers. Conversely, light snow or rain is the most impactful negative predictor, with a coefficient of -0.279, suggesting that such weather conditions are likely to decrease the demand for bike rentals. Lastly, the passing of each year is associated with an increase in

demand, as indicated by the coefficient of 0.233 for 'yr', which may be attributed to the growing popularity or expansion of bike-sharing services.

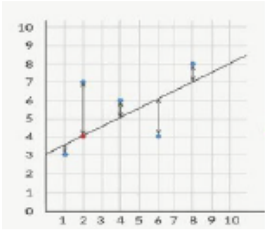
General Subjective Question

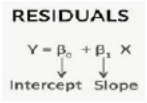
Q1. Explain the linear regression algorithm in detail.

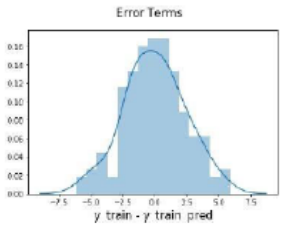
Ans: Linear Regression Model: It is form of predictive modelling technique which describe the relationship between the dependent (Target variable) and independent variables (predictors).

Simple Linear Regression: It is the simplest form of Linear regression, in which we try to find out linear relationship between one dependent and one independent variable.

Multiple Linear Regression: It is the complex form of Linear regression, in which we try to find out linear relationship between one dependent and multiple independent variables.







$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Best Fit Line: Using Linear Regression Algorithm, we try to find the coefficient for independent variable in Best Fit Line which have minimum Residual (Error Term)

Gradient Decent Process: To find best optimized coefficient for independent variables we use Gradient Decent Method.

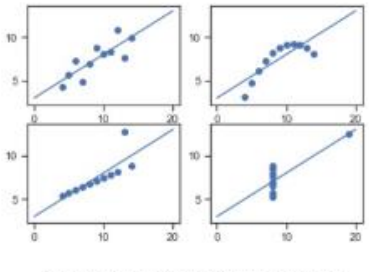
Q2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's Quartet: It is the modal example to demonstrate the importance of data visualization which was developed by the statistician Francis Anscombe in 1973 to signify both the importance of plotting data before analysing it with statistical properties.

Example: Below data set gives an impression that if we do a statistical analysis between x1,y1 or x2,y2 or x3,y3 or x4,y4 then we will get similar kind of infer out of it.

But if we follow the Anscombe's quartet guideline and try to visualize this relationship then it will realize that it is not true.

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89



Graphical Representation of Anscombe's Quartet

Visual Representation of Anscombe's Quartet: It clearly shows that not all four-dataset having linear relationship between x and y variable.

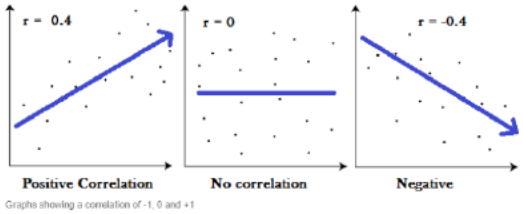
It shows that visual analysis is very important while doing data analysis.

Q3. What is Pearson's R?

Ans: Pearson's R: It is a common correlation coefficient which is used to find a correlation between two variables. It is also known as Pearson's correlation.

Pearson's R values range between -1 and 1

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.



Graphs showing a correlation of -1, 0 and +1

Formula to calculate R (coefficient r) value

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[\,n\sum x^2 - (\sum x)^2\,]\,[\,n\sum y^2 - (\sum y)^2\,]}}$$

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is a technique used in data preprocessing, especially in the context of machine learning and statistics, to standardize the range of independent variables or features of data. Here's a detailed explanation:

1. **What is Scaling?**

   Feature scaling (or scaling of variable values) is a technique to transform the feature values on a common scale. It is part of data preprocessing step in Machine Learning Model building.

2. **Why is Scaling Performed?**

   Scaling helps to bring values of all variables within a specified min/max range, which help algorithm (e.g. Gradient Decent) to design a Model where each variable have equal contribution.

   Feature scaling become important when different variables/independent-variable have values which has a huge different in min/max values for column values.

3. **Difference Between Normalized Scaling and Standardized Scaling:**

   - **Normalized Scaling (Min-Max Scaling):**

- Normalization scales the values into a range of [0, 1] or [-1, 1].
- It’s useful when you need to scale the data to fit into a particular range.
- The formula is (X - X_min) / (X_max - X_min), where X_min and X_max are the minimum and maximum values of the feature respectively.
- Useful when the distribution of the data is unknown or not Gaussian
- Sensitive to outliers
- Retains the shape of the original distribution
- May not preserve the relationships between the data points

- **Standardized Scaling (Z-score Normalization):**
  - Standardization scales data to have a mean of 0 and a standard deviation of 1 (unit variance).
  - It's useful when you want to compare the relative importance of features.
  - The formula is (X - X_mean) / X_std, where X_mean and X_std are the mean and standard deviation of the feature respectively.
  - Useful when the distribution of the data is Gaussian or unknown.
  - Less sensitive to outliers
  - Changes the shape of the original distribution
  - Preserves the relationships between the data points

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans:

VIF (Variance Inflation Factor): Feature scaling (or scaling of variable values) is used to find a correlation (Multicollinearity) between two independent variables.

Value of VIF starts from 1 and has no upper limit.

If VIF is infinite that means there is a strong multicollinearity between those two independent variables, and it is not good for your Model Building

Q6. What is a Q-Q Plot? Explain the Use and Importance of a Q-Q Plot in Linear Regression.

Ans:

he quantile-quantile plot is a graphical method for determining whether two samples of data came from the same population or not.

We can find below data inferences from Q-Q plot:

- Determine whether two samples are from the same population.
- Whether two samples have the same tail
- Whether two samples have the same distribution shape.
- Whether two samples have common location behaviour.

Importance of Q-Q plot

- Since Q-Q plot is like probability plot. So, while comparing two datasets the sample size need not to be equal.
- Since we need to normalize the dataset, so we don't need to care about the dimensions of values.