Assignment-based Subjective Questions

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

Answer:

Based on the analysis of categorical variables in the dataset, we can infer the following effects on the dependent variable (bike rental count):

1. Seasonal Effects:
   - **Winter (coef: 0.086067)** shows a stronger positive effect compared to **Summer (coef: 0.046369)**.
   - This suggests that bike rental demand is relatively higher in winter, which might be due to fewer alternative activities or stable commuting patterns. However, summer also shows an increase, but it is less pronounced.
2. Monthly patterns:
   - **September (coef: 0.075863)** continues to have a strong positive effect on bike rentals, indicating higher demand during late summer and early fall.
   - **July (coef: -0.049569)** shows a reduction in demand, likely due to very hot weather conditions. This indicates higher demand in late summer/early fall, and lower bike-sharing use during harsh weather.
   - This pattern suggests that demand peaks during September, while **Spring (coef: -0.065308)** and extreme summer conditions slightly reduce demand.
3. Weather conditions:
   - **Light Rain/Snow (coef: -0.290409)** has a significant negative impact, reducing rentals, as harsh weather conditions deter users from riding.
   - **Mist (coef: -0.083534)** also has a moderate negative effect, though less severe than light rain or snow.
   - Clear, pleasant weather likely promotes higher rentals, as the negative coefficients reflect a drop in demand when adverse weather is present.
4. Holiday Effect:
   - The inclusion of **Holiday (coef: -0.105739)** shows a notable reduction in bike rentals during holidays
   - This may indicate that people prefer other forms of transportation or are less likely to commute on holidays, hence the drop in demand.
5. Year:

- **Year (coef: 0.233369)** maintains a strong positive effect, indicating a steady growth in bike-sharing demand over time, potentially due to increased adoption of bike-sharing services.

6. Day of the Week Effect:
    - **Sunday (coef: -0.048479)** shows a slight negative impact, indicating that rental demand tends to drop on Sundays compared to weekdays, possibly because of reduced commuting.

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

Answer:

It's important to use **drop_first=True** during dummy variable creation for the following reasons:

1. **Avoid multicollinearity:**
    - When creating dummy variables for a categorical feature, using all categories creates perfect multicollinearity.
    - This can cause issues in some machine learning models, particularly linear models.
2. **Reduce redundancy:**
    - With n categories, you only need n-1 dummy variables to fully represent the information.
    - The dropped category becomes the reference category, and its information is implicitly contained in the other dummies.
3. **Improve model interpretability:**
    - Coefficients for the remaining dummy variables are interpreted relative to the dropped (reference) category.
4. **Prevent the "dummy variable trap":**
    - This trap occurs when the model matrix is not full rank, which can cause issues in model fitting.
5. **Computational efficiency:**
    - Fewer variables mean less computational overhead and potentially faster model training.

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

Answer:

Based on the information provided in the model summary and coefficients, we can infer that among the numerical variables:

Temperature (temp) has the highest correlation with the target variable (bike rental count).

Evidence supporting this:

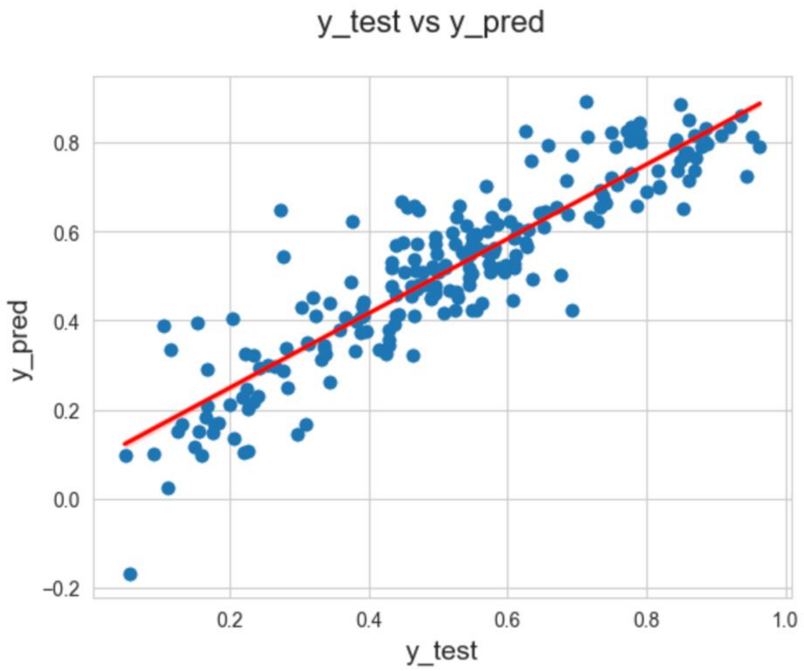
1. **Coefficient value:**
    - Temperature has the highest positive coefficient (**0.491162**) among all variables
2. **Statistical significance:**
    - The p-value for temperature is **0.000**, indicating high statistical significance
3. **Relative importance:**
    - In the table of coefficient values, temperature is at the top, suggesting it has the strongest effect on the target variable

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

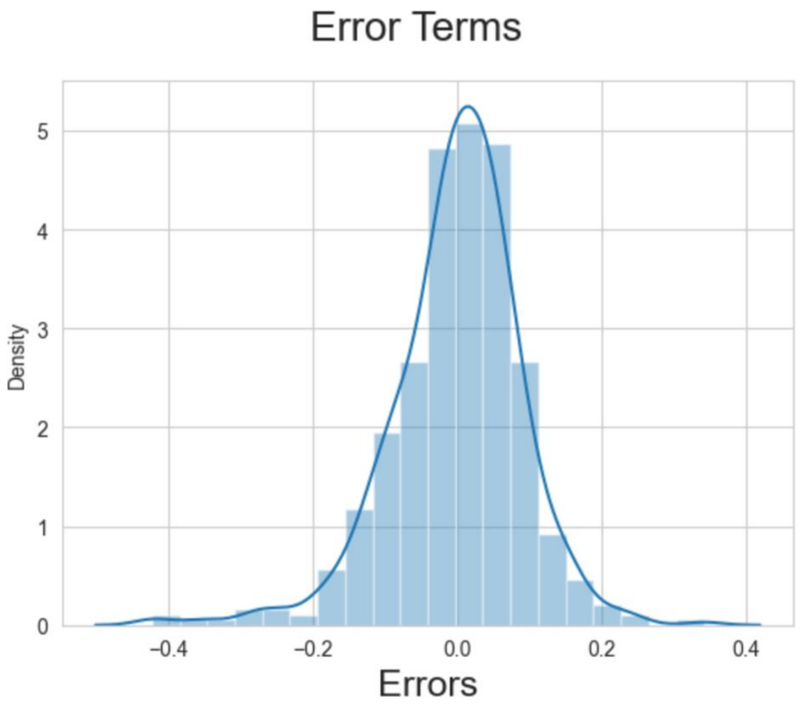
Answer:

To validate the assumptions of Linear Regression after building the model on the training set, we typically follow these steps:

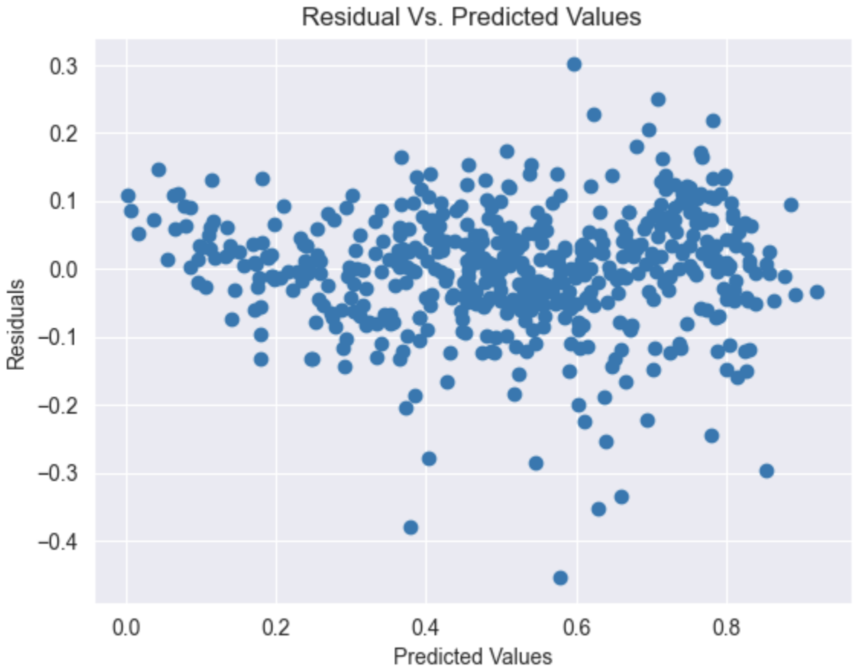
a) **Linear Relationship between dependent and independent variables** – The linearity is validated by looking at the symmetric points distribution around diagonal line of the actual vs predicted plot as shown in below snapshot.



b) **Error terms are normally distributed with mean zero** – Histogram plot helps to understand the normal distribution of error term along with the mean zero. We can clearly depict the same in below screenshot.



c) **Error terms are independent of each other** – There should be no specific pattern observed in error terms with respect to prediction, then only we can say error terms are independent of each other. We see the same in below screenshot.



d) **Error terms have constant variance (homoscedasticity)** – There should be a constant variance in error terms. We can see similar pattern in above screenshot (provided in point a).

e) We used VIF to check for multicollinearity.

f) Also the calculated Mean Squared Error(MSE) (0.0078 for train set), which came very low, suggesting model is a good fit.

g) Also calculated R-squared values (0.8448 for train, 0.8089 for test) which indicates consistent performance.

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

Answer:

Based on the final model results provided, the top 3 features contributing significantly towards explaining the demand of shared bikes are:

1. **Temperature (temp)**
   - Coefficient: **0.491162**
   - Highest positive impact on bike demand
2. **Year (yr)**
   - Coefficient: **0.233369**
   - Second highest positive impact, indicating a strong year-over-year increase in demand.
3. **Winter season**
   - Coefficient: **.086067**
   - Third highest positive impact among the provided features.

General Subjective Questions

1.Explain the linear regression algorithm in detail.

Answer:

Linear regression is a fundamental statistical and machine learning algorithm used for predicting a continuous target variable based on one or more input features. It falls under **supervised learning methods** – in which you have the previous years' data with labels and you use that to build the model. It helps in predicting a dependent variable(Y) based on the given independent variables(x). It is mostly used for finding out the linear relationship between variables and forecasting.

There are two **types** linear regression algorithms –

- **Simple Linear Regression(SLR)** – Single independent variable is used.
  - *Formula:* $Y = \beta_0 + \beta_1 X + \varepsilon$
- **Multiple Liner Regression(MLR)** – Multiple independent variables are used.
  - *Formula:* $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$

  where, β_0 *- value of Y when x=0 (Y-intercept),* ε *– Error Term*

  and β_1 *,* β_2 *,….,* β_n *- Slope or gradient.*

The **strength** of the linear regression model can be assessed using 2 metrics:

- **R^2** or **Coefficient of Determination**
  - Mathematically, it is represented as*:* $R^2 = 1 - (RSS / TSS)$, where: **RSS** = Residual Sum of Squares **TSS** = Total Sum of Squares
- **Residual Standard Error (RSE)**
  - Formula: $RSE = \sqrt{(RSS / (n - p - 1))}$, where **n** = number of observations and **p** = number of predictors (independent variables)

Below are **assumptions** of simple linear regression:

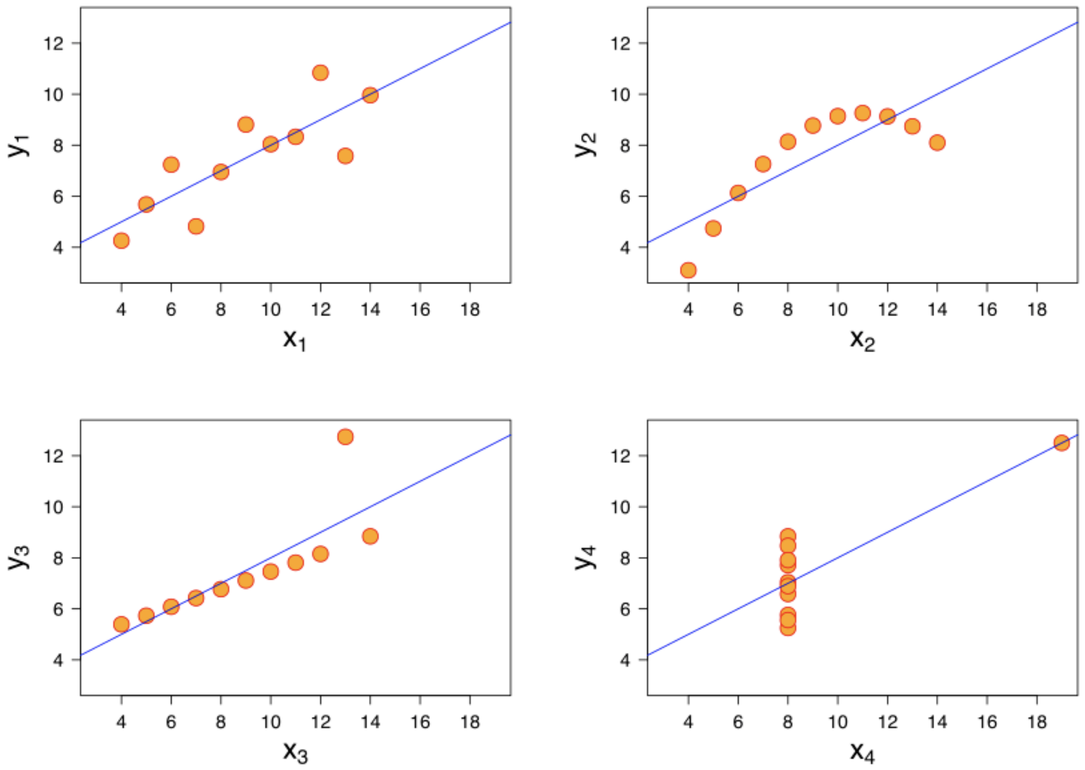
1. Linear relationship between X and Y
2. Error terms are normally distributed (not X, Y)
3. Error terms are independent of each other
4. Error terms have constant variance (homoscedasticity)

Every time we perform a linear regression, we need to test whether the fitted line is significant one or not, for that we need to perform Hypothesis Testing. We also use **p-values** to determine whether a coefficient is significant or not and **F-statistic** to determine whether the overall model fit is significant.

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet comprises four datasets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analysing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough".



It is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties.

3. What is Pearson’s R?

Answer:

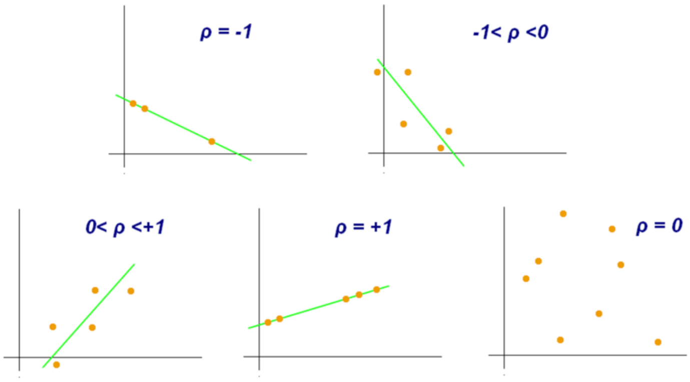
Pearson's R, also known as Pearson's correlation coefficient or simply the correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It measures the degree to which two variables are linearly related.

Although interpretations of the relationship strength vary between disciplines, the table below gives general rules of thumb:

Pearson correlation coefficient (*r*) value	**Strength**	**Direction**
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

Formula: **$R = \Sigma((x - \bar{x})(y - \bar{y})) / \sqrt{(\Sigma(x - \bar{x})^2 * \Sigma(y - \bar{y})^2)}$** where $\bar{x}$ and $\bar{y}$ are the means of x and y respectively.

As you can see in below screenshot, when r is 1 or –1, all the points fall exactly on the line of best fit and when r is greater than .5 or less than –.5, the points are close to the line of best fit. And when r is 0, a line of best fit is not helpful in describing the relationship between the variables



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

Answer:

It's a data pre-processing step applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done, then algorithm only takes magnitude in account, and not units, hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

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.

There are two major methods to scale the variables:

- **Normalization** (Min-Max scaling) - It brings all of the data in the range of 0 and 1.
  - Formula : MinMax Scaling: $x=x-\min(x) / \max(x)-\min(x)$

- **Standardization** (Z-score normalization) – Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**).
  - Formula : Standardisation: $x=x-\text{mean}(x) / \text{sd}(x)$

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

Answer:

VIF calculates how well one independent variable is explained by all the other independent variables combined.

The VIF is given by: VIF_i=1/(1−R_i^2), where '*i*' refers to the i-th variable which is being represented as a linear combination of rest of the independent variables.

If there is a perfect corelation between two independent variables(multicollinearity), then R^2 will be 1, hence VIF will become infinite, as the above formula suggests. The solution to avoid this is to remove or combine highly correlated features.

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

Answer:

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

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.

Few advantages:

1. It can be used with sample sizes also
2. Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

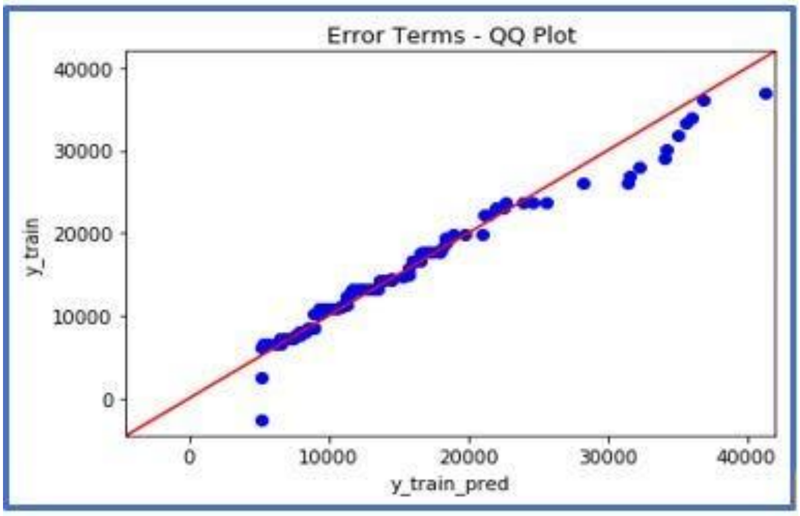
- come from populations with a common distribution
- have common location and scale
- have similar distributional shapes
- have similar tail behaviour

Interpretation:

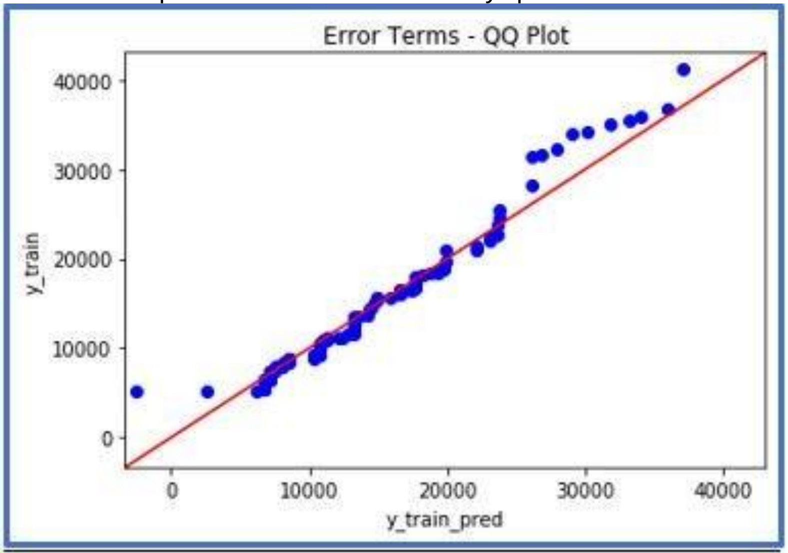
- A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Below are the possible interpretations for two data sets.

a) **Similar distribution**: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
b) **Y-values < X-values:** If y-quantiles are lower than the x-quantiles.



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



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