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?

Ans :

yr : The shift from the first year to the subsequent year (or years) led to a significant growth in demand by **0.234** normalized units, indicating a strong positive trend.

weather_sit with a value 3 mapped to dummy variable ***light_precip*** : Being in a day with light rain/snow/thunderstorm drastically reduces demand by **0.275** normalized units compared to the baseline clear weather.

summer, winter (dummy variable for season) have a small positive impact and ***spring*** has small negative impact in the final model. The variance in bike demand that is shared between ***season*** and ***temp*** is primarily attributed to ***temp*** because it is a continuous, highly predictive variable. Once ***temp*** is in the model, the seasonal dummies are left to explain only the variance that ***temp*** *doesn't* capture.

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

Ans:
We need only K-1 dummy variables which can represent all the combinations. The omitted (dropped) category becomes the baseline (reference) level. the value of the Kth dummy variable can be perfectly predicted by knowing the values of the other K-1 dummies.

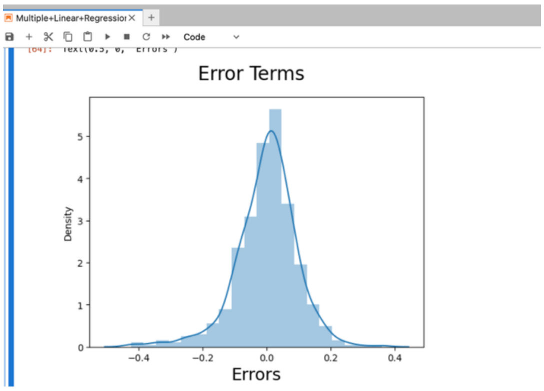
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation
with the target variable? (1 mark) :

Ans :
temp or atemp

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

Ans :

- We validated there is no multicollinearity by ensuring that all our predictor variables are within the acceptable VIF range between >1.0 and <5.0
- We validated the assumption of normal distribution of error terms by plotting a histogram of the error terms (***y_train* – *y_train_cnt*** (predicted value on training set)) and check that the error terms are normally distributed around the mean value of 0 as depicted in the below plot.



```
[64]: Text(0.5, 0, 'Errors')
```

- We can validate the assumption that error terms are independent of each other by looking at Durbin-Watson Statistic , it should be approx ~~ 2.0 , in our case it is 1.98

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

Ans.

Top 3 features are as follows:

1. ***temp*** : temperature in Celsius – Strong Predictor with positive correlation
2. ***yr*** : year (0: 2018, 1:2019) – Moderate Predictor with positive correlation
3. Dummy variable : ***light_precip*** (with the actual value of ***weather_sit*** variable i.e. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) - Moderate predictor with negative correlation.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Ans:

   It is a **linear model** that assumes a linear relationship between the input variables X and the single output variable Y

   The primary goal of the algorithm is to find the **best-fitting straight line** (or hyperplane in multiple dimensions) that minimizes the distance between the line and the actual data points

   Mathematical equation for the Simple Linear Regression is :

   $Y = \beta_0 + \beta_1 X + \in$

   - X: The Independent Variable (predictor).
   - β_0 :The Intercept (the value of Y when X=0).
   - β_1: The Coefficient (the slope) of the predictor, representing the change in Y for a one-unit change in X.
   - $\in$: The Error Term (the residual), representing the difference between the actual Y and the predicted Y.

   Mathematical equation for the Multiple Linear Regression is :

   $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \in$

   Finding the "Best Fit" :

   The algorithm finds the best fit line or optimal coefficients (β values) by minimizing a **cost function**, typically the **Residual Sum of Squares**:

   $RSS = \Sigma(y_i - \hat{y}_i)^2$

   - **y_i** = actual observed value for the i-th data point
   - **$\hat{y}_i$** = predicted value from the regression model for the i-th data point
   - **Σ** = sum across all n data points (i = 1 to n)

   This is done using methods like:

Ordinary Least Squares (OLS): Analytical solution using calculus
Gradient Descent: Iterative optimization approach

Linear regression assumptions:

- Linear relationship between variables
- Independence of observations
- Homoscedasticity (constant variance of errors)
- Normally distributed errors
- No multicollinearity (features aren't highly correlated)

Model Evaluation :

1. R^2 (R-squared / Coefficient of Determination)

Formula: $R^2 = 1 - (RSS/TSS)$

RSS as defined above.
TSS(Total sum of squares): It is the sum of errors of the datapoints from mean of target variable. Mathematically, TSS is:

$$\mathbf{TSS = \Sigma(y_i - \bar{y})^2}$$

Where:

- $\mathbf{y_i}$ = actual observed value for the i-th data point
- $\mathbf{\bar{y}}$ = mean of all observed values ($\bar{y} = (1/n)\Sigma y_i$)
- $\mathbf{\Sigma}$ = sum across all n data points (i = 1 to n)

- Range: 0 to 1 (can be negative for very poor models)
- Interpretation: Proportion of variance in the dependent variable explained by the model
- 0.7-1.0: Strong model
- 0.4-0.7: Moderate model
- Below 0.4: Weak model
  Limitation: R^2 always increases when adding more features, even irrelevant ones

2. Adjusted R^2

Formula: Adjusted $R^2 = 1 - [(1-R^2)(n-1)/(n-p-1)]$

Where n = number of observations, p = number of predictors

- Penalizes adding unnecessary features
- Better for comparing models with different numbers of variables
- Only increases if new variables improve the model more than expected by chance

3. P Values:

- Used to determine the **statistical significance** of individual coefficients

4. F-statistic Value: Tests overall significance of a regression model

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

Answer:

The Anscombe's quartet is a set of four distinct datasets that were created in 1973 by statistician Francis Anscombe. It serves as a classic and powerful demonstration of the importance of graphical analysis in statistics, highlighting why relying solely on summary statistics is misleading.

The Core Insight

All four datasets have nearly identical summary statistics:

- **Mean of x**: 9.0 (for all four)
- **Mean of y**: ~7.5 (for all four)
- **Variance of x**: 11.0 (for all four)
- **Variance of y**: ~4.1 (for all four)
- **Correlation**: 0.816 (for all four)
- **Linear regression line**: y = 3 + 0.5x (for all four)

If you only looked at these numbers, you'd conclude the datasets are essentially the same. But when you plot them, you discover they're radically different.

1. **Dataset I: The Ideal Case**

- **Appearance:** The data points are scattered randomly around the regression line.
- **Inference:** This dataset perfectly satisfies the assumptions of linear regression: the relationship is linear, and the variance of the residuals is constant (**Homoscedasticity**). The OLS model is valid and reliable.

2. **Dataset II: The Non-Linear Case**

- **Appearance:** The data clearly follows a **curvilinear or parabolic pattern**.
- **Inference:** While the linear regression line (Y = 3 + 0.5X) yields the same statistics, it is a poor model. This dataset **violates the assumption of Linearity**. A quadratic or other non-linear regression model would be required for an accurate fit.

3. **Dataset III: The Outlier Case**

- **Appearance:** All data points except one follow a perfect straight line. A single **outlier** significantly pulls the regression line.
- **Inference:** This plot demonstrates the **non-robustness of OLS to outliers**. The single outlier biases the slope of the line. If the outlier were removed, the

R^2 would jump to 1.0, and the coefficient estimates would change dramatically. The OLS results are valid, but they are entirely dependent on that single anomalous point.

4. **Dataset IV: The Leverage Case**

- **Appearance:** All data points are clustered at the same X-value, except for one point that has a high X value.
- **Inference:** This demonstrates a point with **high leverage**. The entire slope of the regression line is determined solely by the position of that single high-leverage point. The remaining points offer no information about the slope. This highlights potential issues with extrapolation and influential observations.

Lessons Learnt:

- Always visualize your data before applying statistical models
- Summary statistics can hide important patterns in the data structure
- Outliers can dramatically affect regression analysis
- Correlation doesn't imply the relationship is linear
- The same model can be appropriate or catastrophically wrong depending on the data's actual structure

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

Ans:

Pearson's R, also known as the Pearson correlation coefficient or Pearson product-moment correlation coefficient (PPMCC), is a descriptive statistic that measures the strength and direction of the linear relationship between two quantitative (continuous) variables.

It is the most common way to measure a linear correlation and is widely used across various fields like data science, finance, and social sciences

The Scale

- **r = +1**: Perfect positive linear relationship (as one variable increases, the other increases proportionally)
- **r = 0:** No linear relationship
- **r = -1:** Perfect negative linear relationship (as one variable increases, the other decreases proportionally)

Values in between indicate the strength:

- 0.7 to 1.0 (or -0.7 to -1.0): Strong correlation
- 0.4 to 0.7 (or -0.4 to -0.7): Moderate correlation
- 0.0 to 0.4 (or 0.0 to -0.4): Weak correlation

What It Actually Measures

Pearson's r measures how well the data points fit a straight line. It's calculated based on how much the variables co-vary relative to their individual variations.

The formula is: $r = \Sigma[(x - \bar{x})(y - \bar{y})] / \sqrt{[\Sigma(x - \bar{x})^2 \times \Sigma(y - \bar{y})^2]}$

In simpler terms: it's the covariance of the two variables divided by the product of their standard deviations.

Critical Limitations

- **Only detects linear relationships**: This is why Anscombe's quartet is so important - Dataset II had a strong relationship but r doesn't capture it well because it's curved, not linear.
- **Sensitive to outliers:** A single extreme point can dramatically change the correlation.
- **Doesn't imply causation**: Two variables can be highly correlated without one causing the other (they might both be caused by a third variable, or the correlation could be coincidental).
- **Assumes continuous data**: Not appropriate for categorical or ordinal data.
- **Can be misleading with non-linear patterns**: You could have a perfect parabolic relationship with r close to 0.

When to Use It

Pearson's R is appropriate when you want to measure linear association between two continuous variables that are roughly normally distributed without extreme outliers. It's one of the most commonly used statistics in research, but like all statistics, it should be accompanied by visualization to ensure you're not missing important patterns in your data.

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

Ans:

Scaling is a data preprocessing technique that transforms numerical features to a common scale or range. It changes the magnitude of the values while preserving the relationships and patterns in the data.

Why is Scaling Performed?

Scaling is essential for several reasons:

- **Algorithm performance**: Many machine learning algorithms (like neural networks, SVM, k-NN, k-means) are sensitive to the magnitude of features. Without scaling, features with larger ranges can dominate the model.
- **Gradient descent optimization**: Algorithms using gradient descent converge much faster when features are on similar scales.
- **Distance-based algorithms**: Methods that rely on distance calculations (k-NN, k-means, PCA) require scaled features to prevent bias toward high-magnitude features.
- **Regularization**: Regularization techniques (L1, L2) penalize coefficients based on magnitude, so unscaled features can be unfairly penalized.
- **Interpretability**: Scaled coefficients are easier to compare when features are on the same scale.

Normalized vs Standardized Scaling

Normalization (Min-Max Scaling)

- **Formula**: X_scaled = (X - X_min) / (X_max - X_min)
- **Range**: Scales data to [0, 1] (or a custom range like [-1, 1])
- **Limitations**: Sensitive to outliers (they can compress the majority of values)

Standardization (Z-score Scaling)

- **Formula**: X_scaled = (X - μ) / σ (where μ = mean, σ = standard deviation)
- **Range**: No fixed range; typically centers around 0 with standard deviation of 1
- **Characteristics**: Preserves the shape of the original distribution

Key Difference:

The fundamental difference is that **normalization** bounds your data to a specific range, while **standardization** centers your data around zero with unit variance but doesn't bound the values to a specific range. Choose based on your algorithm's requirements and whether outliers are present in your data

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

Ans:

VIF (Variance Inflation Factor) can become infinite due to the mathematical formula used to calculate it and the presence of **perfect multicollinearity**.

The VIF Formula

VIF is calculated as:

VIF_i = 1 / (1 - R^2_i)

Where R^2_i is the coefficient of determination when regressing the i-th predictor against all other predictors.

Why Infinity Occurs

VIF becomes infinite when **R^2 = 1**, which happens in cases of **perfect multicollinearity**:

1. Perfect Linear Dependence

When one predictor is an exact linear combination of other predictors:

- Example bike assignment: cnt = casual + registered (all three variables included)

2. Duplicate Features

When the same feature appears twice in the dataset for ex: temp and atemp in bike assignment use case, perhaps with different names.

3. Constant Features

When a feature has zero variance (all values are the same), though this typically causes other issues first.

Mathematical Interpretation

When R^2 = 1:

- The predictor can be **perfectly predicted** by other predictors
- There's **no unique solution** for the regression coefficients
- The predictor adds **no independent information**
- VIF = 1/(1-1) = 1/0 = ∞

Practical Solutions

1. **Remove redundant features**: Drop one of the perfectly correlated variables
2. **Feature engineering**: Combine correlated features into a single meaningful feature
3. **Check for calculation errors**: Sometimes it's a data entry or preprocessing mistake
4. **Domain knowledge**: Use subject matter expertise to decide which variables to keep

VIF Interpretation Guide

- **VIF = 1**: No correlation
- **VIF = 1-5**: Moderate correlation (generally acceptable)
- **VIF = 5-10**: High correlation (investigate)
- **VIF > 10**: Very high correlation (problematic)
- **VIF = ∞**: Perfect multicollinearity (must address)

When you encounter infinite VIF, it's a clear signal that you need to remove or transform features before proceeding with your regression model.

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

Ans:

A **Q-Q plot (Quantile-Quantile plot)** is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, most commonly the normal distribution.

How It Works

- **X-axis**: Theoretical quantiles from the reference distribution (e.g., normal distribution)
- **Y-axis**: Observed quantiles from your actual data
- **Interpretation**: If the data follows the theoretical distribution, points will fall approximately along a straight diagonal line (usually 45-degree line)

The plot compares the quantiles of your data against what those quantiles would be if the data perfectly followed the theoretical distribution.

Use and Importance in Linear Regression

Primary Use: Checking Normality of Residuals

In linear regression, Q-Q plots are primarily used to verify one of the key assumptions: **residuals should be normally distributed**.

Why This Matters

1. **Validity of Inference**:
   - Confidence intervals and hypothesis tests (t-tests, F-tests) assume normally distributed errors
   - Violation affects the reliability of p-values and confidence intervals
2. **Model Reliability**:
   - Non-normal residuals may indicate model misspecification
   - Could suggest missing variables, wrong functional form, or outliers
3. **Prediction Accuracy**:
   - Severely non-normal residuals can affect prediction intervals
   - May indicate heteroscedasticity or other issues

Interpreting Q-Q Plots in Regression

Patterns and What They Mean

1. **Points on the diagonal line**:
   - ✓ Residuals are normally distributed
   - Model assumption is satisfied
2. **S-shaped curve**:

   - Heavy tails (more extreme values than normal)
   - May indicate outliers or a distribution with higher kurtosis
3. **Inverted S-shape**:
   - Light tails (fewer extreme values)
   - Lower kurtosis than normal distribution
4. **Points curving above the line at both ends**:
   - Right-skewed distribution
   - Consider log transformation of the response variable
5. **Points curving below the line at both ends**:
   - Left-skewed distribution
   - May need transformation
6. **Systematic departure from the line**:
   - Clear non-normality
   - Investigate model specification or consider transformations

When to Use Q-Q Plots

- **After fitting a regression model**: Check residuals, not the raw response variable
- **During model diagnostics**: Along with other diagnostic plots (residual vs. fitted, scale-location)
- **Before making inferences**: Especially important for small sample sizes
- **When violations are suspected**: Based on other diagnostic tests

Limitations

1. **Subjective interpretation**: What counts as "close enough" to the line can be debatable
2. **Sample size dependent**: Small samples naturally show more deviation from the line
3. **Not a formal test**: Complement with statistical tests like Shapiro-Wilk if needed
4. **Only checks one assumption**: Must use other diagnostic tools for complete model validation

Practical Considerations

- **Minor deviations are acceptable**: Perfect normality is rare in real data
- **Focus on severe violations**: Especially at the tails with many observations
- **Robust methods available**: If normality is severely violated, consider robust regression or transformation
- **Large samples**: Central Limit Theorem provides some protection; normality assumption becomes less critical

The Q-Q plot is an essential diagnostic tool that provides visual, intuitive insight into whether your regression model's residuals meet the normality assumption, helping you build more reliable and valid statistical models.