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? (3 marks)
   Answer:
   a. Season vs count: season shows significant pattern in usage of boombikes. Spring and summer have reported high usage of bikes.
   b. Year vs count: it demonstrates that over a period usage of bikes has increased.
   c. Month vs count: it demonstrates and re-affirm the observation of season that month between May – October when weather is clean and warm usage of bikes is higher.
   d. Holiday vs count: it demonstrates that usage of bikes increases and decrease during holidays. This is further re-affirm during EDA where it is observed that usage of bike among casual users on Saturday and Sunday is significantly high.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
   Answer: Dummy variables are created from single dataset and hence will have interdependency resulting in multicollinearity causing impact in linear regression. If we retain all variable in regression model, then intercept will have perfect multicollinearity. To prevent this we drop first variable in order to simply the model without losing any information and coefficient created from linear regression are interpretable and statistically valid.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   Answer: highest correlation observed from pair-plot is between 'cnt' vs 'atemp'

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   Answer: Assumption of linear regression is validated by residual analysis by evaluating "error term" in which we evaluate if error term is normally distributed. This is done to ensure homoscedasticity in data variable i.e residual have constant variation at every level of predictor variable.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
   Answer:
   a. **Impact of weather conditions on boombikes revenue:** variables such as temperature, humidity and wind speed have strong influence on boombikes customer to bikes.
   b. **Impact of time:** Attributes associated with time ex: day, weekday, working day, month, season etc. have significant impact on usage of bike by customer of boombikes. Time and weather to-gether have visible and distinct impact on company business.
   c. **Impact of events:** Additional attributes of holiday bring additional revenue from casual users of boombikes. From the data it is visible that usage of bikes by casual users is highest during holidays. While registered users provides steady revenue durign workding and weekdays, causal users bring good influx of revenuw during holiday.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   Answer: In simple terms it is statistical approach by which programming skills is utilised to read and interpret the data for any corelation. Conceptually linear regression is used to learn the relationship between a dependent variable and one or more independent variables. Mathematically it is explained as model with goal to find the best-fitting linear relationship which minimizes the difference between actual values vs predicted values.
1. **Linear Equation:** Simple linear regression with one independent variable is explain by
   $y=\beta 0+\beta 1x+\epsilon$
   where:
   (y) is the dependent variable.
   (x) is the independent variable.
   ($\beta 0$) is the intercept of the line.
   ($\beta 1$) is the slope of the line.
   (ϵ) is the error term (the difference between the observed and predicted values).

   For multiple linear regression the equation is:

   $y=\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta nxn+\epsilon$

2. In linear regression coefficient ($\beta 0, \beta 1,\ldots,\beta n$) are estimated to minimize sum of squared difference between observed values (y) and predicted value (y^)
3. Ordinary Least Squares (OLS): It is widely used approach for estimating coefficient.
4. Steps followed for linear regression evaluation:
   a. **Data Collection:** obtaining data containing dependent and independent variable.
   b. Data Preparation: it includes
      i. Data cleaning
      ii. Train – test split
      iii. Scaling the data as necessary.
   c. **Model Specification:** define dependent and independent variable.
   d. **Fitting the model:** Estimate coefficient ($\beta 0,\beta 1,\ldots,\beta n$) using OLS method. This involves solving the following normal equation:
      $\beta=(X^TX)-1X^Ty$
      where X is the matrix of input features, y is the vector of observed values, and β is the vector of coefficients.
   e. **Evaluating model:** it can be done using following approach
      i. R-squared (R^2)
      ii. Mean Squared Error (MSE)
      iii. Root Mean Squared Error (RMSE)
   5. **Interpretation:** Statistical tests (like t-tests) is usually used to determine significance of coefficient.
   6. **Prediction:** Developed model is fitted to make prediction on new data.

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

Answer: Anscombe's quartet was created by statistician Francis Anscombe to demonstrate importance of graphing data when analysing it and its effect of outliers and other influential observation. Anscombe's quartet is set of four dataset which have almost identical descriptive statistics but have different distribution when graphed. Each dataset consists of eleven (x, y) points.

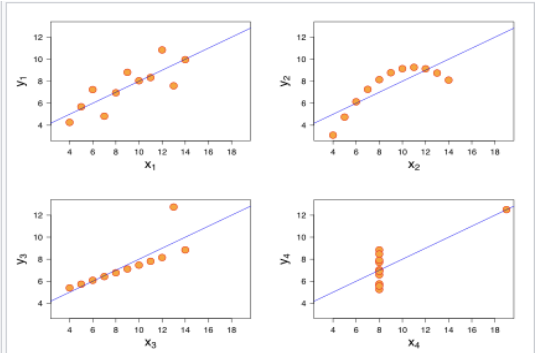
- **The datasets are as follows.**

Dataset I		**Dataset II**		**Dataset III**		**Dataset IV**	
x	**y**	**x**	**y**	**x**	**y**	**x**	**y**
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

- **Statistical description of all datasets is as below:**

Property	**Value**	**Accuracy**
Mean of x	9	exact
Sample variance of x: s^2_x	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y: s^2_y	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	y = 3.00 + 0.500x	to 2 & 3 decimal, respectively
Coefficient of linear regression:	0.67	to 2 decimal places

- **Graphical representation of all four dataset**



The four datasets composing Anscombe's quartet. All four sets have identical statistical parameters, but the graphs show them to be considerably different

- The first scatter plot is for simple linear relationship, corresponding to 2 correlated variables, where *y* could be modelled as gaussian with mean linearly dependent on *x*.

- In second graph it is obvious, that relation between 2 variable is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.

- In the third graph, the modelled relationship is linear, but should have a different regression line. The calculated regression is offset by the one outlier, which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph is an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

3. What is Pearson's R? (3 marks)
   Answer: It was developed by Karl Pearson from a related idea introduced by Francis Galton in the 1880s, and for which the mathematical formula was derived and published by Auguste Bravais in 1844. The naming of the coefficient is thus an example of Stigler's Law. Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean of the product of the mean-adjusted random variables; hence the modifier product-moment in the name

   The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Pearson correlation coefficient (r)	Correlation type	Interpretation	Example
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the same direction.	Baby length & weight: The longer the baby, the heavier their weight.
0	No correlation	There is no relationship between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the opposite direction.	Elevation & air pressure: The higher the elevation, the lower the air pressure.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)?
   Answer: Scaling is used in data preprocessing and transforming the values of variables in a dataset to a similar scale. This is often done to ensure that the variables contribute equally to the analysis and modelling process, especially when they have different scales or units. Scaling can help improve the performance of certain machine learning algorithms, particularly those sensitive to the scale of the input features, such as gradient descent-based algorithms.
   - Normalized scaling, known as min-max scaling, involves scaling the values of a variable to a range between 0 and 1. This is typically done by subtracting the minimum value of the variable and then dividing by the difference between the maximum and minimum values.
   - Standardized scaling involves transforming the values of a variable so that they have a mean of 0 and a standard deviation of 1. This is typically done by subtracting the mean of the variable and then dividing by the standard deviation.
   - The key difference between normalized scaling and standardized scaling lies in how the values are transformed. Normalized scaling adjusts the values to a specific range, while standardized scaling adjusts them to have a specific mean and standard deviation.

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

   Answer: The Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in regression analysis. Multicollinearity occurs when one predictor variable in a regression model can be linearly predicted from the others with a substantial degree of accuracy. VIF quantifies how much the variance of a regression coefficient is inflated due to multicollinearity. A VIF value becomes infinite when R^2_i is equal to 1. This situation occurs when there is perfect multicollinearity, meaning that the i^{th} predictor can be perfectly predicted by a linear combination of the other predictors. In such cases, the denominator of the VIF formula ($1 - R^2_i$) becomes zero, causing the VIF to approach infinity.

$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2}$$

   VIF become infinite because of following reasons:

   - Perfect Multicollinearity: when one predictor is an exact linear function of one or more other predictors perfect multicollinearity occurs.
   - Duplicate Variables: Including the same variable more than once creates perfect multicollinearity.
   - Linear Combinations: If a predictor is a linear combination of other predictors, this will also result in perfect multicollinearity.

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

   Answer: A Q-Q plot, is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, often the normal distribution. In the context of linear regression, Q-Q plots are particularly useful for evaluating the assumption that the residuals (errors) of the regression model are normally distributed.

   - **Plotting Procedure:**
     - **Quantiles**: For example, the median is the 0.5 quantile.
     - **Theoretical Quantiles**: These are the quantiles of the theoretical distribution which are compared with data against evaluation (e.g., normal distribution).
     - **Sample Quantiles**: These are the quantiles of the sample data.
   - **Creating a Q-Q Plot**:
     - Sorting the data and determine its sample quantiles.
     - Calculating the corresponding theoretical quantiles from the specified theoretical distribution.
     - Plotting the sample quantiles against the theoretical quantiles.
   - **Interpretation**:
     - **Straight Line**: If the data follows the theoretical distribution, the points will approximately lie on a straight line (the 45-degree reference line).
     - **Deviations**: Systematic deviations from the line suggest departures from the theoretical distribution.
   - **Importance of a Q-Q Plot in Linear Regression**: In linear regression, one of the key assumptions is that the residuals (errors) are normally distributed. This assumption is critical because many statistical tests and confidence intervals rely on it.
   - **Use and Importance:**
     - **Assessing Normality**:
       1. **Residual Analysis**: By plotting the residuals of a regression model it can be determine if residuals follow normal distributions.

2. **Model Validation**: If the residuals deviate significantly from the normal line, it suggests that the model assumptions might be violated, which can affect the validity of the model's statistical tests and confidence intervals.

- **Detecting Outliers and Extreme Points**:
- **Identifying Skewness and Kurtosis**:
  1. **Skewness**: If the points form an S-shaped curve, it indicates skewness in the residuals.
  2. **Kurtosis**: If the points are concave up or down relative to the line, it indicates issues with the kurtosis
- **Improving Model Fit**: **Model Diagnostics**: Using a Q-Q plot to diagnose non-normality in residuals can guide transformations of the response variable or adding/removing predictors to improve model fit.