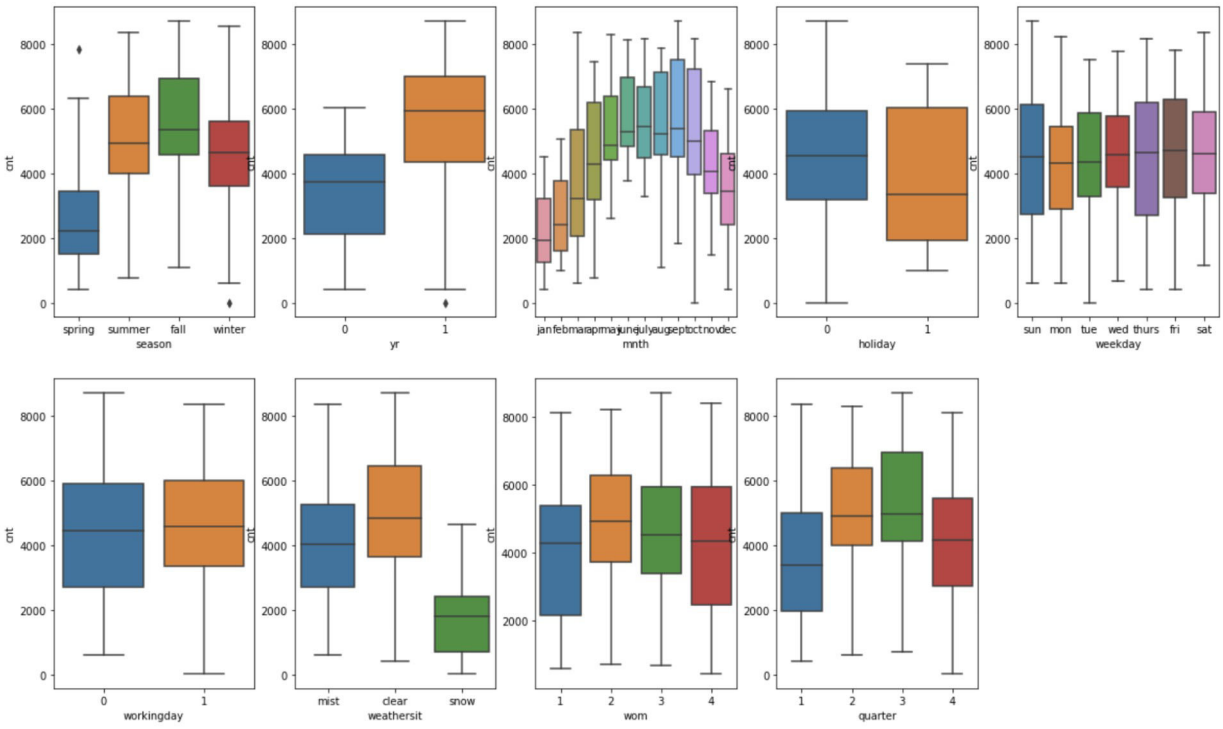
Assignment-based Subjective Questions

Q1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans:

1. **Season**: cnt keeps on increasing from spring to fall and decreases in winter with highest cnt in fall(Autumn).
2. **Year**: There is significant growth in cnt from year 2018 to 2019.
3. **Month**: cnt keeps on increasing from January to July and then starts declining till December.
4. **Holiday**: cnt is significantly less on holidays compared to non holidays.
5. **Weekday**: There is not significant difference between days for cnt.
6. **Woking Day**: cnt is almost same for working and non working day.
7. **Weather**: cnt is highest on clear day, then comes misty day and lowest on snowy day.
8. **Week Of Month**: Second week of month is carrying higher demand than rest of the weeks.
9. **Quarter of Year**: Quarter 2 and 3 carry highest demand then 4 and last 1, representing pattern similar to seasons.



Q2: Why is it important to use **drop_first=True** during dummy variable creation?

Ans: Pandas get_dummies function on drop_first=False does not give a base state and create variables equal to number categories in the categorical variable, introducing an extra variable while drop_first=True creates a base state represented by all 0s for one of the category and therefore creating n-1 variables for representing same information. Thus preventing correlation between dummy variables; drop_first=True can represent same information in lower number of variables. This will also have some o impact of adj. R^2, for drop_first=True it will higher compared to drop_first=False.

For example: Representing Low, Medium and High
drop_first=False: Low (1,0,0); Medium (0,1,0); High (0,0,1) , sum of 3 vars is always 1
drop_first=True: Low (0,0); Medium (1,0); High (0,1)

Q3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: **atemp or temp** shows maximum correlation with cnt [Not including instant, casual and registered in pair-plots]

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

Ans: Running model summary and checking the following params:

1. Linear Relationship between X and Y.
    1. **Feature Coefficients**: variable coefficients are not 0 representing that is their impact on target variables.
    2. **Feature p-values**: Feature p-values are less than 5% showing their high significance in reflecting outcome of dependent variable.
    3. **F-Statistic and Prob(F-Statistic)**: High F-Statistic and low Prob(F-Statistic) represent that model has not fitted just by chance and their actual relation between independent and dependent variables.
    4. **R^2 and Adj. R^2**: These signify the model accuracy on training data that is amount of variance in training data that is captured by model. High values indicate greater accuracy. For Adj. R^2 its in relation with number of features along with model accuracy.
2. **Residual Analysis**:
    1. Normal distribution of residuals by distribution plot of errors.
    2. Scatter plot of errors is randomly distributed, not representing any patterns.
    3. Homoscedasticity, by plotting scatter plot of residuals with y predicted to check constant variance of error terms.

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

Ans: Max 3 from modulus of Coefficients of features are:

Number	Feature	Coefficient
1	atemp	0.4593
2	yr	0.2344
3	weathersit_snow	-0.1740

General Subjective Questions

Q1: Explain the linear regression algorithm in detail.

Ans: Linear Regression is supervised machine learning technique used to predict a continuous numeric variable dependent on other feature variables. Feature variable can be numeric or categorical but for processing all variable are given a numerical representation for training and testing.

Considering there is linear relationship between independent and dependent variable, mathematically it can be represented as
y is dependent variables
x_1 , x_2 , …. x_n are independent variables
θ_1 , θ_2 , …. θ_n are coefficient of independent variables respectively
c is constant
ε is error

$y = \theta_1 x_1 + \theta_2 x_2 + \ldots. + \theta_n x_n + c + \varepsilon$

1. Load and identification of numeric and categoric variables
2. Categoric variables are converted to numeric with dummy variables.
3. Splitting data to Training and Test data in ratio of (70%, 30%) or (80%, 20%) respectively.
4. Normalization or standardization of training data.
5. Feeding the training data to the algorithm with actual label(dependent var) and features as independent vars.

Training the model – Algorithm fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ and c values.

Cost Function (J):
By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the θ and c values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).

$J = MIN([\Sigma (pred_i - y_i)^2]/n)$ where n is row count of training data.

Cost function(J) of Linear Regression is the **Root Mean Squared Error (RMSE)** between predicted y value (pred) and true y value (y)

Gradient Descent:
To update θ and c values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line the model uses Gradient Descent. The idea is to start with random θ and c values and then iteratively updating the values, reaching minimum cost.

Q2: Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet explains importance of data visualization where data may represent separate patterns or have outliers can represent same statistical measures. Represented in Fig. 1 data can actually represent the same statistics of mean, variance, correlation and regression line but having completely different representations on visualization.

Visualization becomes an important tool in actually understanding data. Visualising the data before applying various algorithms to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help to identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.

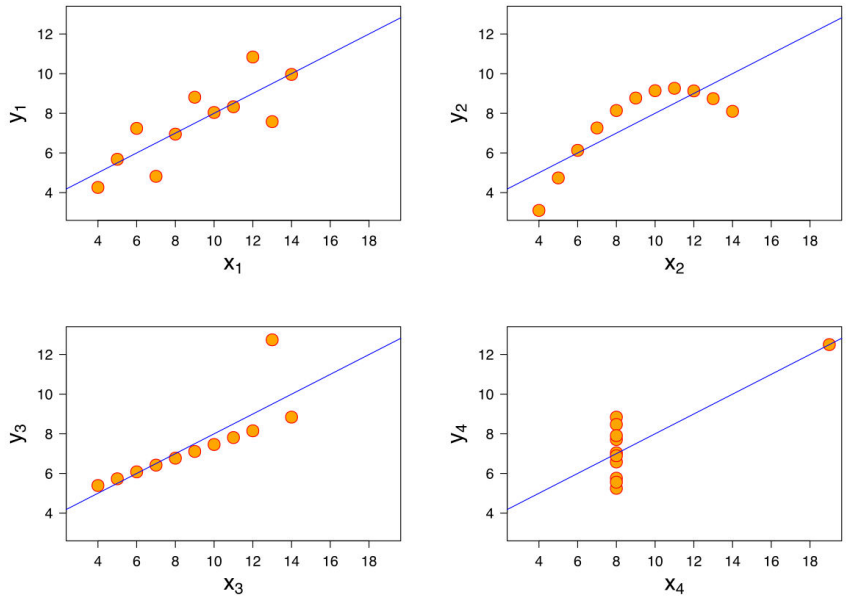


Fig. 1

Q3: What is Pearson's R?

Ans: **Pearson correlation coefficient** (**PCC)** also known as **Pearson's *r***, the Pearson product-moment correlation coefficient (PPMCC), the bivariate correlation is a measure of a linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviation thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

The measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation. For example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).

Formula for r_{xy} by substituting estimates of the covariances and variances based on a sample into the formula above. Given paired data $\{ (x_1 , y_1) , \dots , (x_n , y_n)\}$:

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum x_i^2 - n\bar{x}^2}\ \sqrt{\sum y_i^2 - n\bar{y}^2}}.$$

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

Ans: **Scaling:** Step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Reason: 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.
Scaling only effects the coefficient of independent variables.

Normalization: It brings all of the data in the range of 0 and 1. Even the outliers are brought in the range 0 and 1 thus lose information about outliers. Formula:

normalized(x) = (x - min(x)) / (max(x) – min(x))

Standardization: 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**). It does not lose information about the outliers. Formula:

standardized(x) = (x -μ_x)/ σ_x

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

Ans: VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. VIF is calculated for features excluding the label to check the relation between a feature with rest of the features, perfect correlation will have VIF = ∞.

VIF = 1 / (1 – R^2)

Q6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans: Q-Q(quantile-quantile) plots play a very vital role to graphically analyse and compare two probability distributions by plotting their quantiles against each other. If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x.

Use: It’s very important to know whether the distribution is normal or not so as to apply various statistical measures on the data and interpret it in much more human-understandable visualization. Q-Q helps to check if the data is distribution normally.

Significance:

To fit a linear regression model, check if the points lie approximately on the line, and if they don't residuals and errors are not Gaussian. Implying that estimator is not Gaussian either, so the standard confidence intervals and significance tests are invalid.

Helps in a case 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.

So Q-Q plot can check whether data is apt for linear regression or not.