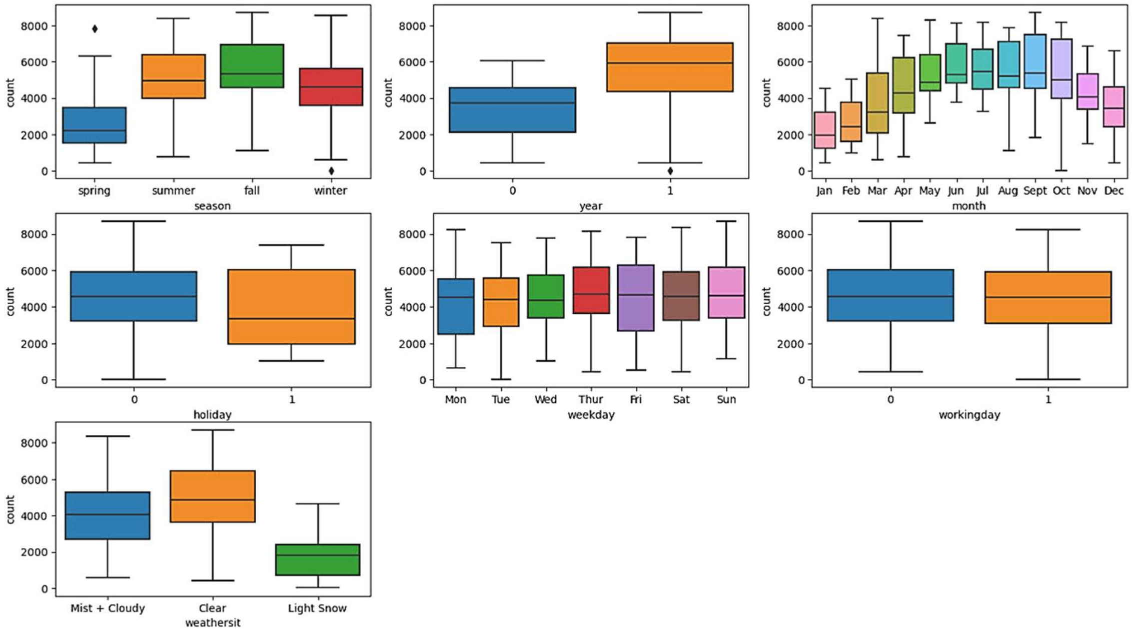
Assignment-based Subjective Questions

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



Inferences

a) **Season** : Customer count varies across the seasons, Season 3 'Fall' appears to have highest median count and season 1 'Spring' has lowest customer count

b) **Year**: The 2019 (yr = 1) has significantly higher customer counts compared to the 2018 (yr = 0).

c) **Month**: The monthly trend suggests a peak in customer counts during summer months (e.g., Jun, July and Aug), with a noticeable drop during colder months (e.g., Jan, Feb, and Dec).

d) **Holiday:** The number of customers is lower on holidays (holiday = 1) as compared to non-holidays (holiday = 0).

e) **Weekday**: There doesn't appear to be a significant variation in customer counts across weekdays. The counts are fairly consistent.

f) **Working day**: Customer counts are slightly higher on working days (working day = 1) than on non-working days.

g) **Weather situation**: Better weather conditions correlate with higher customer counts. Poor weather (e.g., weathersit indicating bad conditions) shows a significant drop in customer numbers.

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

Dummy variable is created in pandas using pd.get_dummies (), it is important to use drop_first=True to:

- **Avoid dummy variable trap:**

  a) The **dummy variable trap** occurs when one or more dummy variables are perfectly correlated with each other (multicollinearity).

  b) If all categories of a categorical variable are encoded as dummy variables, their values will always sum to 1, creating redundancy in the data. This redundancy can confuse machine learning algorithms and lead to unreliable model coefficients.

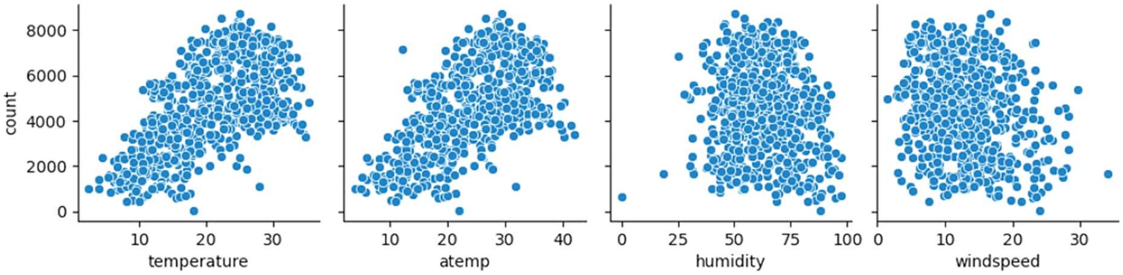
- **Reduce number of features:** This simplification can make model more computational efficient
- **Improve model interpretability:**

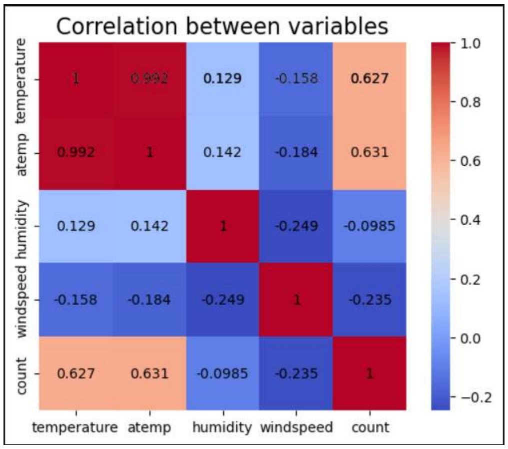
  a) With drop_first=True, the dropped category becomes the reference category.

  b) The coefficients of the remaining dummy variables indicate their impact relative to this reference category.

Dummy variable created for Bike Sharing assignment were for weather-situation, season, month and weekday.

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



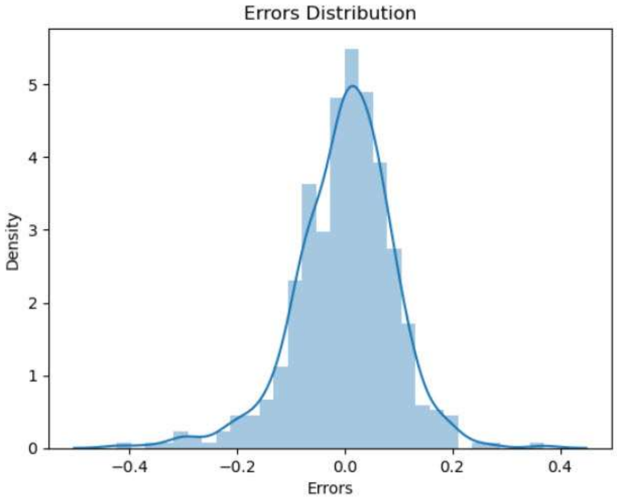


Looking at the pair-plot and correlation matrix, atemp (0.0631) has the highest correlation with target variable count followed by temperature (0.627).

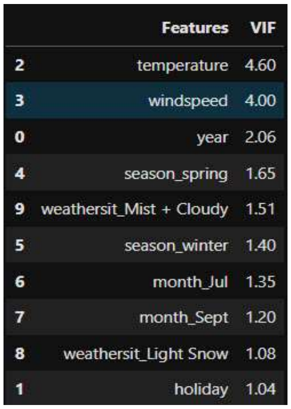
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

a) There should be linear relationship between Independent variables and dependent variable. Numeric variables are visualised in previous question to check the linear relationship between predictors and target variable.

b) So we can see that the residual distribution is normal and centred around zero. Hence our assumption for Linear Regression is valid



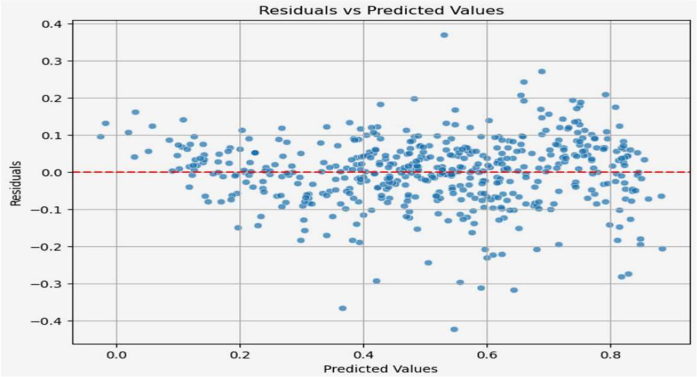
c) Linear Regression assumes there is little or **no multicollinearity** in data. The variables should be independent of each other i.e. no correlation should be there between the independent variables. To check the assumption, we used a correlation matrix or VIF score. Our final model has VIF score for all the Features less than 5



	Features	VIF
2	temperature	4.60
3	windspeed	4.00
0	year	2.06
4	season_spring	1.65
9	weathersit_Mist + Cloudy	1.51
5	season_winter	1.40
6	month_Jul	1.35
7	month_Sept	1.20
8	weathersit_Light Snow	1.08
1	holiday	1.04

d) **No Autocorrelation**: The error terms should be independent of each other. Autocorrelation can be tested using the Durbin Watson test. **Durbin-Watson** value of final model is 2.002, which signifies there is no autocorrelation.

e) **Homoscedasticity**: The variance of the error terms should be constant i.e the spread of residuals should be constant for all values of X. This assumption can be checked by plotting a residual plot. If the assumption is violated then the points will form a funnel shape otherwise they will be constant. Since the points are randomly scattered with no pattern, the model satisfies the assumptions of linearity and homoscedasticity.



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

The top 3 features contributing significantly towards explaining the demand of the shared bikes are:

a) Temperature: A coefficient value of ‘0.451455’ indicated that a unit increase in temperature variable increases the bike hire numbers by 0. 451455 units.

b) Weathesit_Light Snow: A coefficient value of ‘-0.286408 indicated that a unit increase in Weathesit_Light Snow variable decreases the bike hire numbers by 0. 234092 units.

c) Year: A coefficient value of ‘0.234092 indicated that a unit increase in year variable increases the bike hire numbers by 0. 234092 units.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Linear regression is a supervised learning algorithm used for predicting a continuous target variable based on one or predictor variables. It assumes a linear relationship between the independent variables (X) and the dependent variable (y). The model predicts (y) by fitting a straight line (or hyperplane in higher dimensions) through the data points.

Linear Regression is of two types: Simple and Multiple

Simple linear regression has only independent variable and model has to find the linear relationship of it with the dependent variable

Whereas, In Multiple Linear Regression there are more than one independent variables for the model to find the relationship.

Equation of simple Linear Regression:

y = b0 + b1X

y= Dependent variable
b0 = intercept
b1 = coefficient or slope
X is independent variable

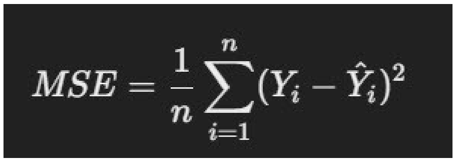
Equation of Multiple Linear Regression:

y = b0 + b1x1 + b2x2 + b3x3……., bnxn

where b1,b2,b3,bn are the coefficient of independent variable x1,x2,x3….,xn and y = dependent variable

Main aim of the Linear Regression model is to find the best fit line and the optimal values of intercept and coefficients that minimises the residuals.
Residuals are the difference between actual and predicted value

Linear Regression uses MSE (mean squared error) as the cost function:



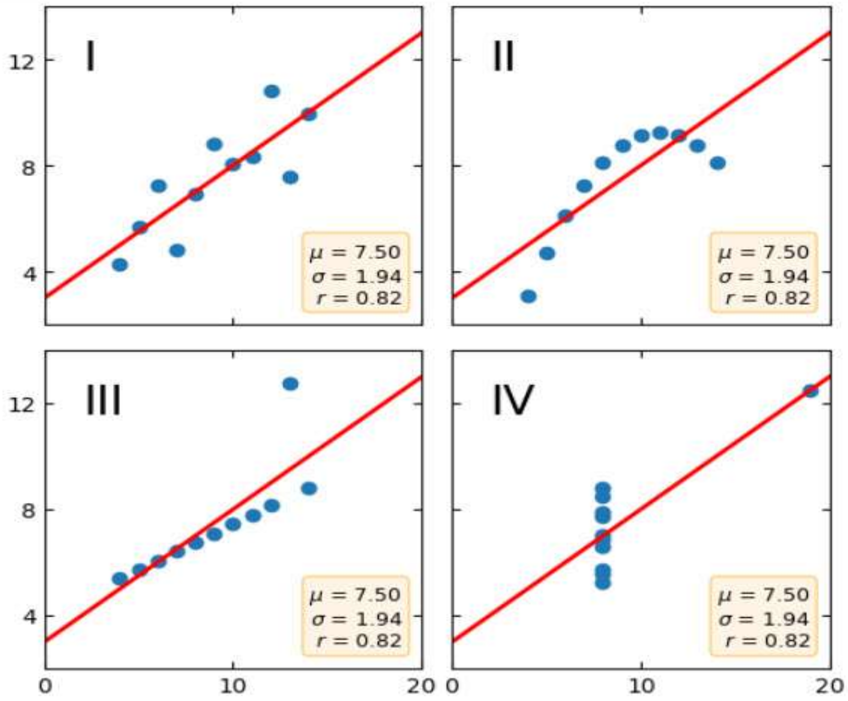
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Assumptions of Linear Regression:

a) Linearity : Linear relationship between X and y
b) Independence : Observations are independent of each other
c) Homoscedasticity : Variance of residuals are constant across all the levels of X
d) Normality : Residuals are normally distributed

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet is a set of four datasets that demonstrate the importance of visualizing data before drawing conclusions based on statistical summaries. The quartet, created by statistician Francis Anscombe in 1973, shows that datasets with nearly identical statistical properties (mean, variance, correlation, regression line, etc.) can have drastically different distributions and visual relationships.



Interpretation of the four datasets:

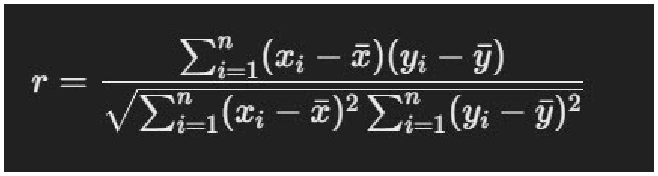
a) First dataset: A simple linear relationship with minor scatter around the regression line.
b) Second Dataset: A curvilinear relationship that does not fit a straight-line model, even though the regression line exists.
c) Third Dataset: A linear relationship influenced by a single outlier. Without the outlier, the trend would be much different.
d) Fourth Dataset: Most of the data points have the same X value, with one significant outlier driving the linear relationship.

Key takeaways:

Visualization of data is necessary prior to predictive or statistical modelling, as it helps to detect any outliers, non-linearity and other patterns that may affect model performance.

3. What is Pearson's R?

Pearson Correlation Coefficient is the statistical measure that quantifies the strength and direction of linear relationship between two continuous variables. The value of r ranges between -1 and +1, -1 being perfect negative relationship and +1 being perfect positive relationship and 0 means no linear relationship.



$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where,

(r): Pearson correlation coefficient

(xi, yi): Individual data points

(x bar, y bar): Mean of variables x and y

n: number of data points

Uses include **Correlation Analysis**, **Feature Selection** (identification of correlated parameters, **Hypothesis testing** (test of the existence of linear relationship between two continuous variables, **Predictive modelling** (feasibility of predictive modelling based on strength of correlation)

Pearson's R captures only linear relationships and may fail to identify non-linear patterns.

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

In statistical modelling and machine learning, scaling is the data pre-processing step, which transforms the feature of dataset to a specific range or distribution, ensuring all the features contributes equally to the model's performance. Scaling is done because:

Model sensitivity: Our data may contain features with different magnitude, unit and range and if scaling is not performed than algorithm tends to weigh high value magnitude and ignore the other features which will result in faulty modelling or biased weight updates.

Improving Performance: Scaling also makes the model simple and speed up calculations in algorithm. Scaling leads to faster convergence of optimization algorithms.

Interpretability: Scaling can make the coefficients of model like Linear Regression more interpretable, as each feature contributes proportionately.

Normalised scaling and standardized scaling are the commonly used techniques used to scale the data.

Difference between Normalized Scaling and Standardized Scaling:

a) **Normalized Scaling** also know min-max scaling. Transforms the data to a fixed range ([0, 1]) whereas **Standardized Scaling** also known as (Z-score) centres the data around zero with 1 variance.

b) **Normalized Scaling** is highly sensitive to outliers on the contrary Standardized is less sensitive to outliers.

c) **Normalized Scaling** is used when we don't know about the distribution whereas standardized is used when the distribution is normal.

d) **Normalized Scaling** Formula: (X-min(X)/max(X)-min(X)), **Standardized Scaling** Formula: (X- mean/standard deviation)

e) **Normalized Scaling** is useful for algorithm k-NN, SVM, and Neural Networks and **Standardized Scaling** Suitable for PCA, regression, and any algorithm sensitive to distribution.

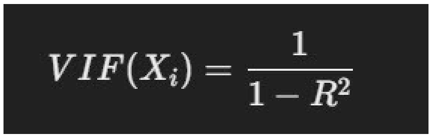
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Variance Inflation Factor shows the relationship between independent variables, measures the degree of multicollinearity in a regression model. When VIF becomes infinite it typically indicates one or more predictors are linear combination of others (perfect multicollinearity), makes model unstable and coefficients unreliable.

Causes of multicollinearity may be:

a) Due to the presence of duplicate columns, means columns having identical or near identical values(e.g., X1=X2)
b) Derived features: one predictor is the linear transformation of other(e.g., X2 = 2* X1+3)
c) Failing to drop one level of categorical variable when encoding it as dummy variable.

Formula



$$VIF(X_i) = \frac{1}{1 - R^2}$$

Here, R2 is Coefficient of determination when Xi is regressed on all other predictors. Perfect multicollinearity can be handled by removing redundant variables, checking the correlation matrix (high correlation closer to 1 and -1) among independent variables and fixing dummy variable trap.

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

A Q-Q (Quantile-Quantile) plot is a graphical tool, used to see if a dataset follows a particular theoretical distribution like normal, exponential or uniform. It is widely used to assess whether the data (or residuals, in the context of regression) follows a specific distribution.

In linear regression, residuals (errors) are the differences between observed and predicted values. A Q-Q plot is used to check whether these residuals follow a normal distribution, which is an assumption in linear regression.

Importance and use in Linear Regression:

a) Linear Regression assumes residuals are normally distributed. This assumption can be validated by Q-Q plot thus ensuring Hypothesis testing and reliable parameter estimation
b) It also helps in outlier detection
c) Model Diagnostics: Deviation from normality indicates that either transformation is needed or model is not well specified.
d) The validity of t-tests, F-tests and confidence intervals relies on normal residuals
e) Improving prediction: As normal distribution leads to reliable and better prediction