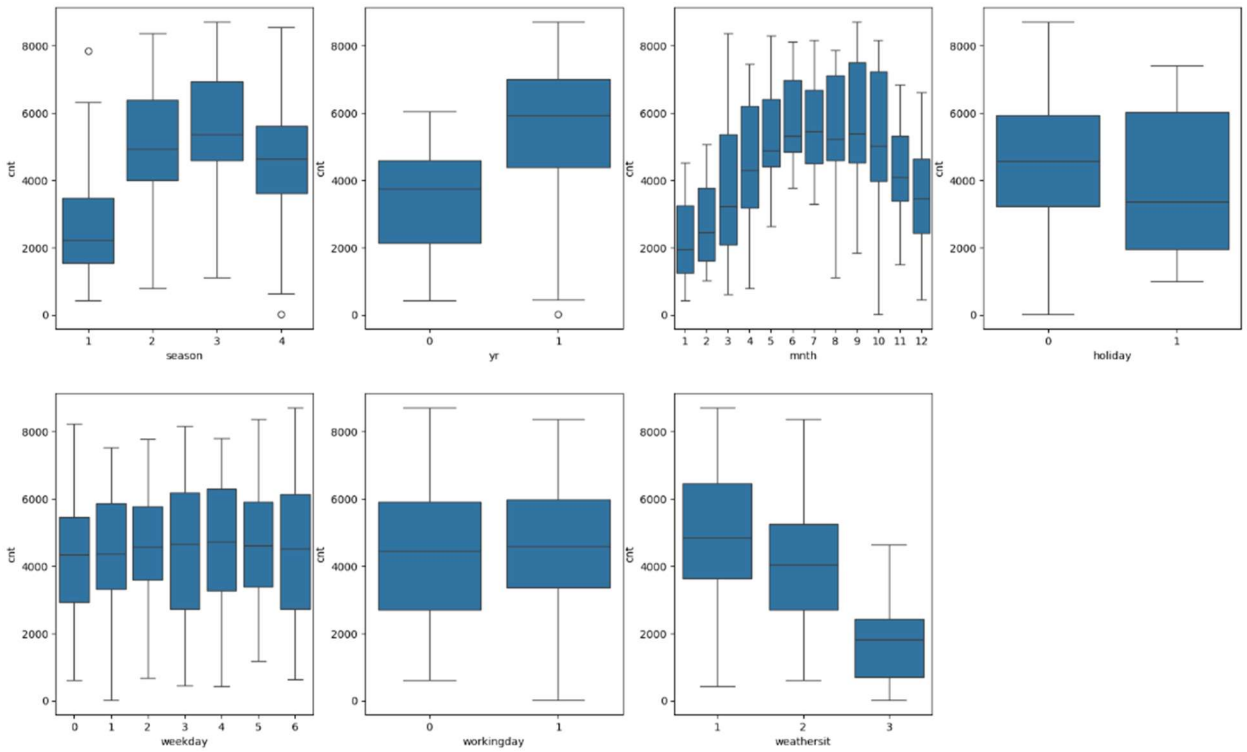
Assignment-based Subjective Questions

Q1) From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)



Answer : Observations of Categorical Variable:

a) If we observe during seasons (1: spring, 2: summer, 3: fall, 4: winter) during summer (2 & 3) the ride count is high

b) 2019 the rides were more as it gets popularity over time from the previous year (2018). This would infer that the next year (2020) the company can expect more rides compared to 2019.

c) Months also shows similar pattern like season as during summer and holiday months like May, June and July rides were more and during winter and snow period there are less rides

d) During Holidays the rides were less and during working days the rides were more

e) During weekday does not shows any significant changes in the pattern

f) Similarly working day rides were less, as it is already shows in holiday

g) “Weather sit” is plays important role, as we have observed like in season and month, Clear sky contributes more rides compared to cloudy and snow.

Q2) Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

Answer : We use "**drop_first=True"** to prevent multi collinearity.

When we create dummy variables for a categorical feature, we introduce new columns, one for each category. However, these columns are not entirely independent.

There's a perfect linear relationship between them. This is known as multicollinearity.

Multicollinearity occurs when independent variables in a regression model are highly correlated. This can lead to unstable estimates, inflated standard errors, and difficulty in interpreting the model.

By setting drop_first=True when creating dummy variables, we drop the first category, effectively making it the baseline. This prevents the perfect linear relationship between the columns, thereby avoiding multi collinearity.

From our data set :

Consider a categorical variable seasons (1: spring, 2: summer, 3: fall, 4: winter) with 4 categories: Without drop_first=True, we would create three dummy variables:

seasons _ spring

seasons _ summer

seasons _ fall

seasons _ winter

If a data point belongs to the spring, then seasons _ spring

will be 1, and the other three will be 0. This creates a perfect linear relationship:

seasons _ spring= 1 - seasons _ summer- seasons _ fall- seasons _ winter

By setting drop_first=True, you drop one category (e.g., seasons _ spring). Now, if a data point is not seasons _ summer (or) seasons _ fall (or) seasons _ winter then it must be seasons _ spring. The information about the dropped category is implicitly captured in the remaining columns.

Points to Ponder:

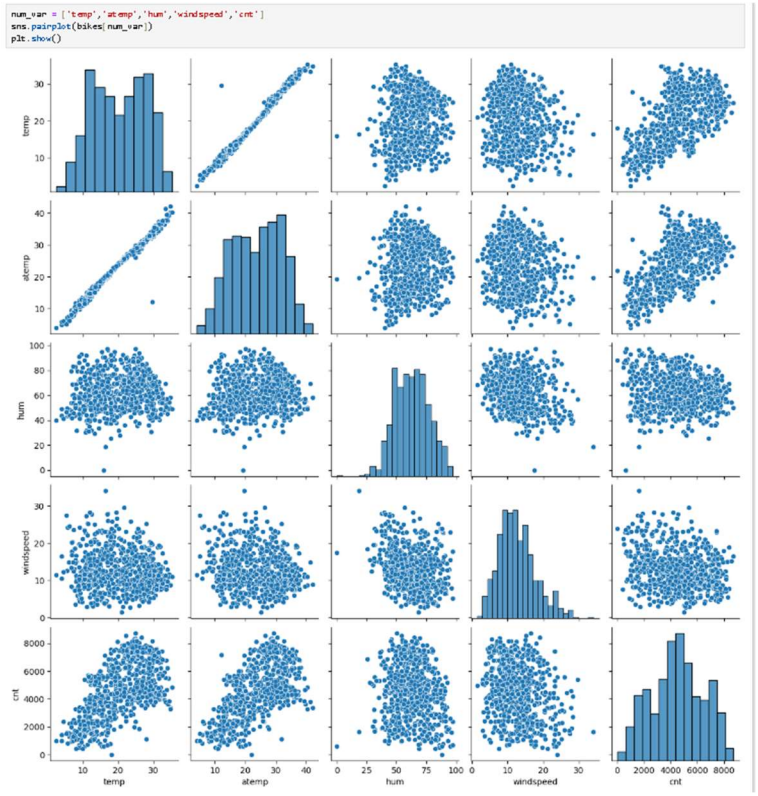
Using drop_first=True reduces the number of dummy variables by one.

1) It prevents multicollinearity.
2) It makes the model interpretation easier.
3) This ensure that your model is more robust and reliable.

Q3). Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer : By looking at the pair-plot “temp” or “atemp” has the highest correlation with the target variable

```
num_var = ['temp','atemp','hum','windspeed','cnt']
sns.pairplot(bikes[num_var])
plt.show()
```



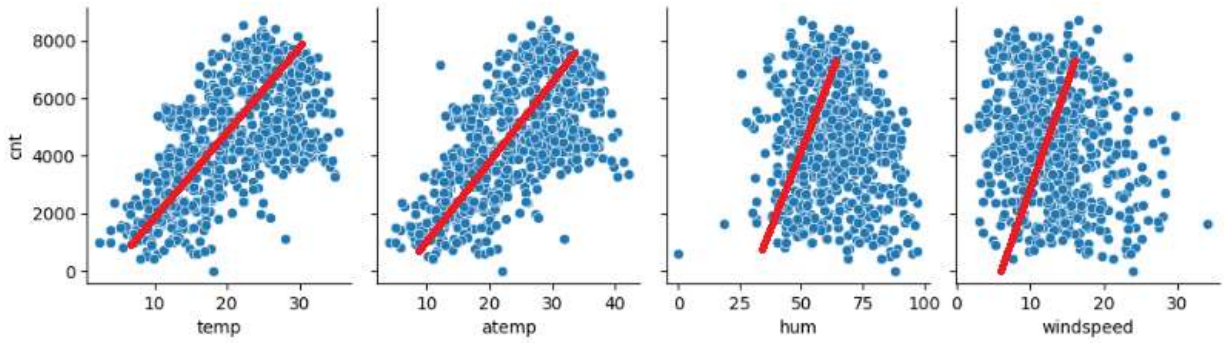
Q4). How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer :

1) Linearity between input features (X_i) and output variable(Y)

   There should be linear relationship between input features and output variable

   This can be observed during EDA by sc**atter plots.** Visualize the relationship between the dependent variable and each independent variable. A linear pattern suggests linearity. By observing the below scatter plot we could conclude that the below numerical vairbales (temp, hum and windspeed) are showing linear pattern.



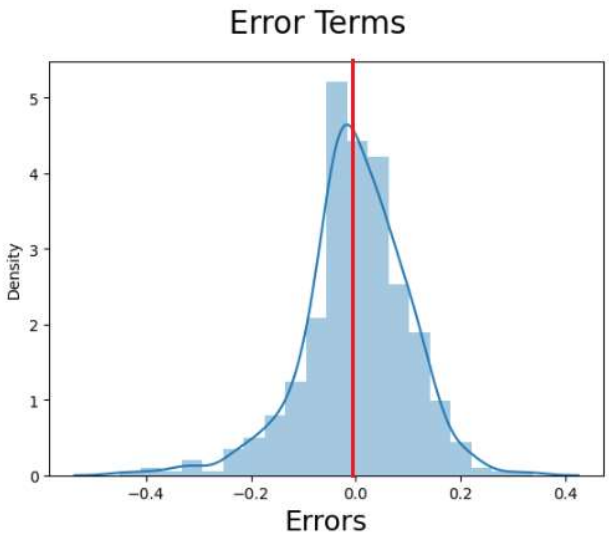
2) Error terms are normally distributed.

   If we observe the below histogram of error terms they are normally distributed and shows bell shape curve. This confirms the good validity of the predicted model.

```
y_train_cnt = lr_3.predict(X_train_lm)
```

```
# Plot the histogram of the error terms
fig = plt.figure()
sns.distplot((y_train - y_train_cnt), bins = 20)
fig.suptitle('Error Terms', fontsize = 20)                  # Plot heading
plt.xlabel('Errors', fontsize = 18)                         # X-Label
```

```
Text(0.5, 0, 'Errors')
```

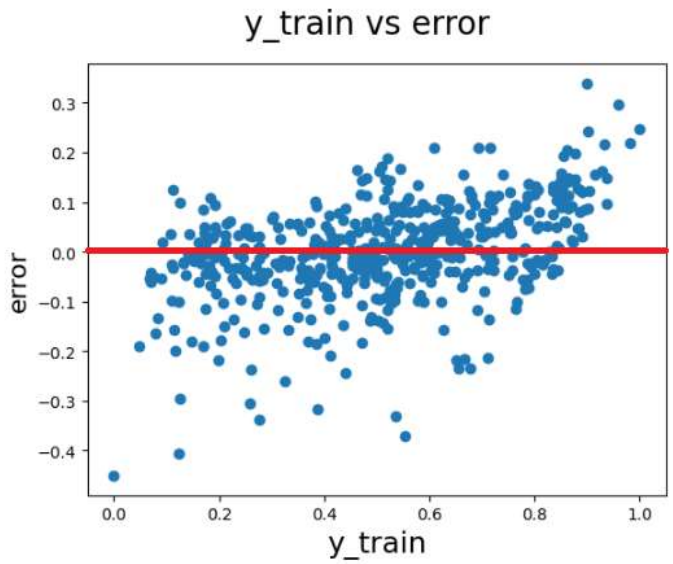


3) Error terms are independent of each other :

If we compute the errors of the model, they are well distributed along the mean and there are no observed pattern in this plot so this model is well suited for linear regression.

```
fig = plt.figure()
error =  y_train-y_train_cnt
plt.scatter(y_train, error)
fig.suptitle('y_train vs error', fontsize = 20)              # Plot heading
plt.xlabel('y_train', fontsize = 18)                         # X-Label
plt.ylabel('error', fontsize = 16)
```

```
Text(0, 0.5, 'error')
```



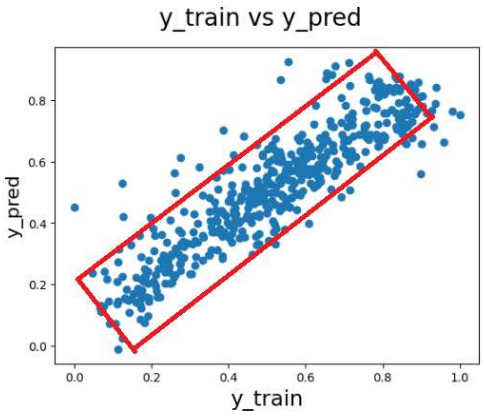
4) Error terms have constant variance (**homoscedasticity**):

The variance should not increase (or decrease) as the error values change.

Also, the variance should not follow any pattern as the error terms change.

```
[66]: fig = plt.figure()
      plt.scatter(y_train, y_train_cnt)
      fig.suptitle('y_train vs y_pred', fontsize = 20)              # Plot heading
      plt.xlabel('y_train', fontsize = 18)                         # X-Label
      plt.ylabel('y_pred', fontsize = 16)
```

```
[66]: Text(0, 0.5, 'y_pred')
```



Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer : From RFE fit my top 3 features contributing significantly towards explaining the demand of the shared bikes are ‘temp’, ‘weathersit’ and ‘yr’.

As per RFE on the final model :

```
rfe = RFE(lm)            # running RFE
rfe = rfe.fit(X_train, y_train)
list(zip(X_train.columns,rfe.support_,rfe.ranking_))
[('season', False, 2),
 ('yr', True, 1),
 ('holiday', False, 3),
 ('weekday', False, 4),
 ('workingday', False, 5),
 ('weathersit', True, 1),
 ('temp', True, 1),
 ('windspeed', True, 1)]
```

General Subjective Questions

Q1) Explain the linear regression algorithm in detail. (4 marks)

Answer : Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables.

- Dependent variable: The variable we want to predict (often denoted as Y).
- Independent variables: The variables used to make predictions (often denoted as X).
- Linear relationship: The relationship between the dependent and independent variables is represented by a straight line.

A simple linear regression model with one independent variable can be represented as:

$Y = \beta0 + \beta1*X + \varepsilon$

Where:

- Y is the dependent variable
- X is the independent variable
- β0 is the intercept (the value of Y when X is 0)
- β1 is the coefficient of X (the change in Y for a one-unit change in X)

For multiple independent variables, the equation becomes:

$Y = \beta0 + \beta1*X1 + \beta2*X2 + ... + \beta n*Xn + \varepsilon$

Where:

- X1, X2, ..., Xn are the independent variables
- β1, β2, ..., βn are their respective coefficients

Model Training:

The goal of linear regression is to find the best values for the coefficients (β0, β1, β2, ...) that minimize the error between the predicted values and the actual values. This is typically done using the ordinary least squares (OLS) method, which minimizes the sum of the squared residuals (differences between predicted and actual values).

Evaluation Metrics:

- Mean Squared Error (MSE): The average of the squared differences between the predicted and actual values.
- Root Mean Squared Error (RMSE): The square root of the MSE, providing a measure of error in the same units as the dependent variable.
- R-squared: A statistical measure that represents the proportion of variance in the dependent variable that is explained by the independent variables.

Assumptions of Linear Regression:

- Linearity: The relationship between the variables is linear.
- Normality : Error terms are normally distributed
- Independence: The observations are independent of each other.
- Homoscedasticity: The variance of the error term is constant for all values of X.

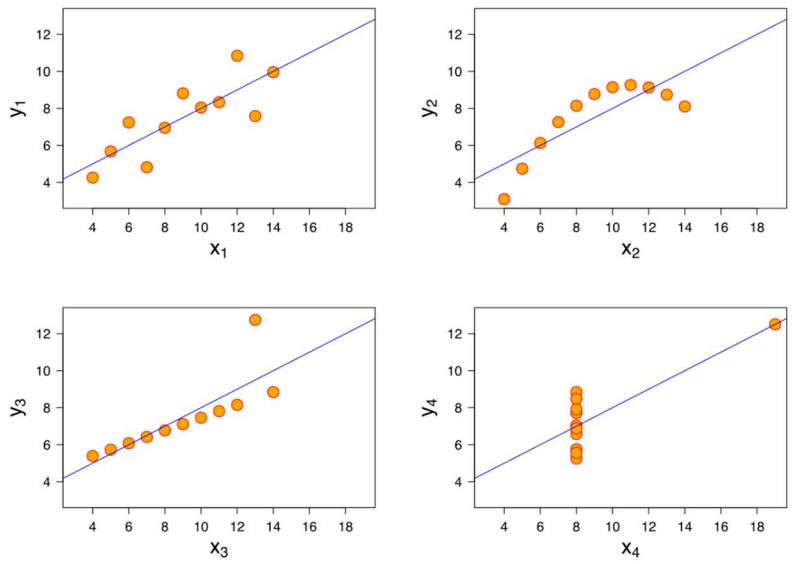
Q2). Explain the Anscombe's quartet in detail. (3 marks)

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 analyzing 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".

For all four datasets:

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x: s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y: s_y^2	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 and 3 decimal places, respectively
Coefficient of determination of the linear regression: R^2	0.67	to 2 decimal places



Q3). What is Pearson's R? (3 marks)

Answer :

Pearson's Correlation Coefficient (r)

Pearson's r is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. It's a value between -1 and 1.

Key characteristics:

Range: -1 to 1

Interpretation:

-1: Perfect negative correlation (as one variable increases, the other decreases)

0: No correlation (no linear relationship)

1: Perfect positive correlation (as one variable increases, the other increases)

Measures linear relationship:

It only assesses linear relationships, not non-linear ones.

Affected by outliers:

Outliers can significantly impact the value of r.

Usage:

When we want to measure the strength of a linear relationship between two continuous variables.

When the data is normally distributed.

Limitations:

Only measures linear relationships.

Sensitive to outliers.

Q4). What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer :

Scaling :

Scaling is a data preprocessing technique that involves transforming numerical features into a common scale. This is essential because many machine learning algorithms, especially distance-based algorithms are sensitive to the scale of features. For example the linear regression each feature is ranges with wide range of boundary of values, to avoid this numerical instability we need to do scaling.

Why is Scaling Performed :

Improves Algorithm Performance:

Most machine learning algorithms assume features are on a similar scale. Without scaling, features with larger values can dominate the learning process, leading to suboptimal results.

Speeds Up Computation:

Some algorithms converge faster when features are scaled.

Better Interpretation:

Scaled features can make it easier to compare and interpret coefficients in models like linear regression.

1) Min-Max Scaling(Normalized Scaling) :

Rescales features to a specific range, typically 0 to 1.

Sensitive to outliers.

Formula :

X_scaled = (X - X_min) / (X_max - X_min)

Where:

X_scaled is the scaled value

X is the original value

X_min is the minimum value in the feature

X_max is the maximum value in the feature

2) Standardized Scaling (Z-score Scaling)

Rescales features to have a mean of 0 and a standard deviation of 1.

Less sensitive to outliers than normalization.

Formula:

X_scaled = (X - mean) / standard_deviation

Where:

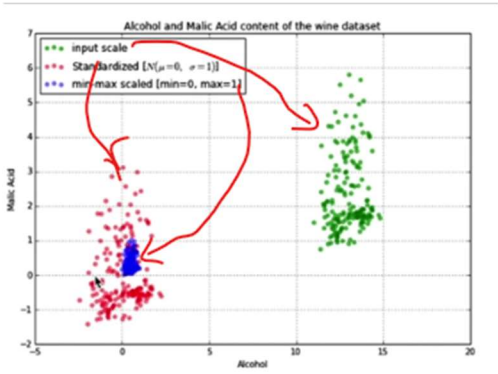
X_scaled is the scaled value

X is the original value

mean is the mean of the feature

standard_deviation is the standard deviation of the feature

Difference between Min-Max Scaling (normalized) scaling and standardized scaling :



Normalization Min-Max scaling (0 to 1) :

It is always between (0 to 1). It always better as it takes care of outliers. As all the outliers are kept in either 0 (or) 1 When you don't know the exact range of your data and outliers are a concern then use Min-Max Scaling.

Standardization (mean-0, sigma-1): When you know the exact range of your data when we are sure that the outliers are captured with in the range and then we can use Standardization.

Q5). You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer :

VIF (Variance Inflation Factor) is a measure of the influence of multi collinearity on the variance of the estimated regression coefficients. Essentially, it quantifies how much the variance of a coefficient is inflated due to the presence of other correlated predictors.

When VIF is Infinite :

A VIF value of infinity indicates perfect multi collinearity.

This means that one independent variable is an exact linear combination of one or more other independent variables.

In simpler terms, there's a perfect relationship between these variables.

Causes of Perfect Multi collinearity:

Duplicate Variables: Having the same variable included twice in the model.

Linear Combinations: One variable being a linear function of another (e.g., weight in kilograms and pounds).

Dummy Variables Trap: Including all levels of a categorical variable without omitting one as a reference category.

Consequences of Perfect Multicollinearity:

The model becomes unsolvable.

The regression coefficients cannot be estimated.

The VIF calculation breaks down, resulting in an infinite value.

Detection and Solutions:

Correlation Matrix: Examine the correlation matrix for values close to 1 or -1.

VIF Calculation: Calculate VIF for each independent variable. An infinite value indicates perfect multicollinearity.

Remove Redundant Variables: Eliminate one of the perfectly correlated variables.

Combine Variables: If the variables are closely related, consider creating a composite variable.

Center and Scale Variables: While this doesn't directly address perfect multicollinearity, it can sometimes improve numerical stability and reduce the impact of multicollinearity.

In essence, an infinite VIF is a clear signal of a serious issue in your dataset, and it's crucial to address it before proceeding with model building.

Q6). What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
(3 marks)

Answer :

Q-Q Plot :

A Q-Q (Quantile-Quantile) plot is a graphical method used to assess if a set of data follows a particular probability distribution.

It compares the quantiles of two probability distributions.

Use and Importance of a Q-Q Plot in Linear Regression

1) Assessing Normality of Residuals

one of the key assumptions is that the residuals (the differences between observed and predicted values) are normally distributed.

A Q-Q plot of the residuals can help assess this assumption:

Straight Line: If the residuals are normally distributed, the points will lie approximately along a 45-degree line.

S-Shaped Curve: Indicates heavy tails (leptokurtic) or light tails (platykurtic).

Bowed Out or In: Suggests skewness in the residuals.

2) Detecting Outliers

Outliers in the residuals can be identified as points that deviate significantly from the straight line in the Q-Q plot.

3) Checking Model Fit

A Q-Q plot can provide insight into how well the regression model fits the data. If the model fits poorly, the residuals may not follow a normal distribution, and this will be evident in the Q-Q plot.