Assignment-based Subjective Questions

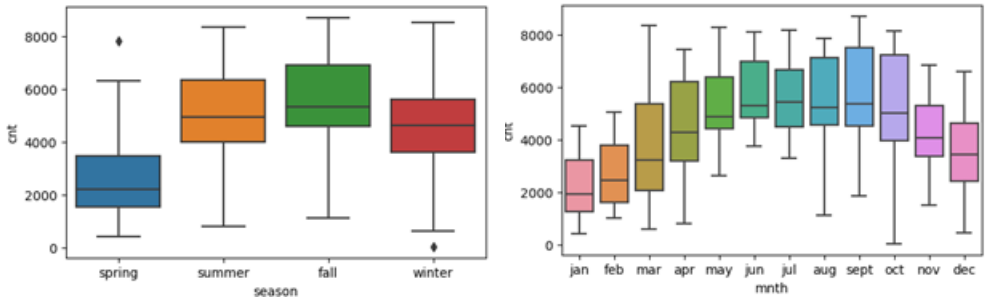
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer:

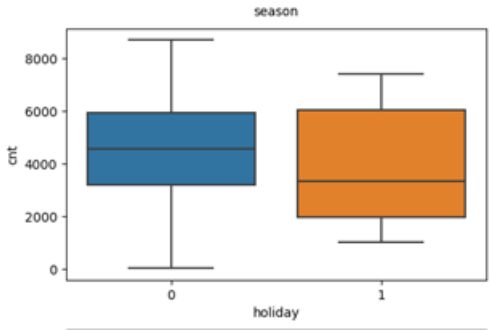
We have 'season', 'yr' ,'mnth' , 'holiday' , 'weekday' , 'workingday 'and 'weathersit' as categorical variable in the given 'day' dataset for bike sharing system assignment.

After doing EDA on categorical variable, initially we can infer that:

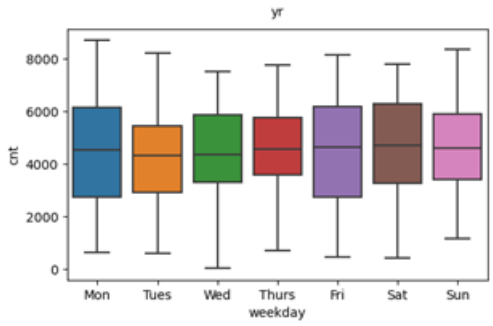
Based on the examination of categorical variables in the dataset, it can be deduced that bike rental rates exhibit an upward trend during the summer and fall seasons, particularly in the months of September and October.



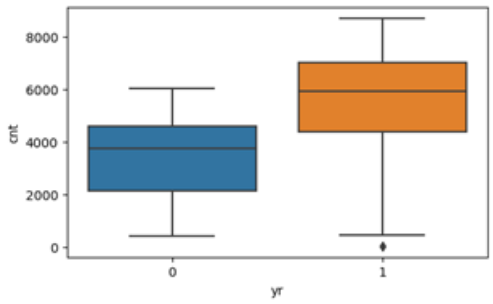
A holiday variable have affected the mean demand negatively



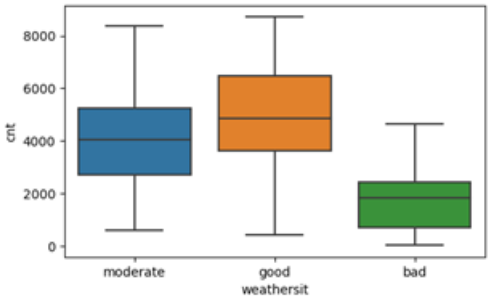
No proper inference about bike demand can be drawn from Weekday.



Furthermore. The year 2019 appears to correlate with higher bike rental rates compare to previous year 2018.



Bike demand is elevated during clear weather conditions and reaches its lowest point during bad weather.



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

   Answer :

   Using drop_first=True when creating dummy variables is important for several reasons:

   Mitigating Multicollinearity: By excluding one dummy variable, we prevent perfect multicollinearity issues, which can destabilize the regression model.

   Enhancing Interpretability: The inclusion of all dummy variables can complicate the interpretation of coefficients. Dropping one dummy variable facilitates a clearer understanding, as the coefficients of the remaining variables represent changes relative to the omitted category.

   Optimizing Efficiency: Including unnecessary dummy variables can impose computational overhead without providing additional insights. Dropping one dummy variable improves computational efficiency and resource utilization.

   In essence, employing drop_first=True ensures a more stable, interpretable, and efficient linear regression model, particularly when dealing with categorical variables.

   Imagine we have two categories: "Category A" and "Category B."

   Without dropping the first category, the encoding would be:

   Category A: 0

   Category B: 1

   But with drop_first=True, it drops Category A, and the encoding becomes:

Category B: 1

Here, if the variable is 0 for "Category B," it means the observation is associated with "Category A." So, drop_first=True helps keep the necessary information about the categories in a more streamlined way for regression analysis, avoiding issues of redundancy.

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

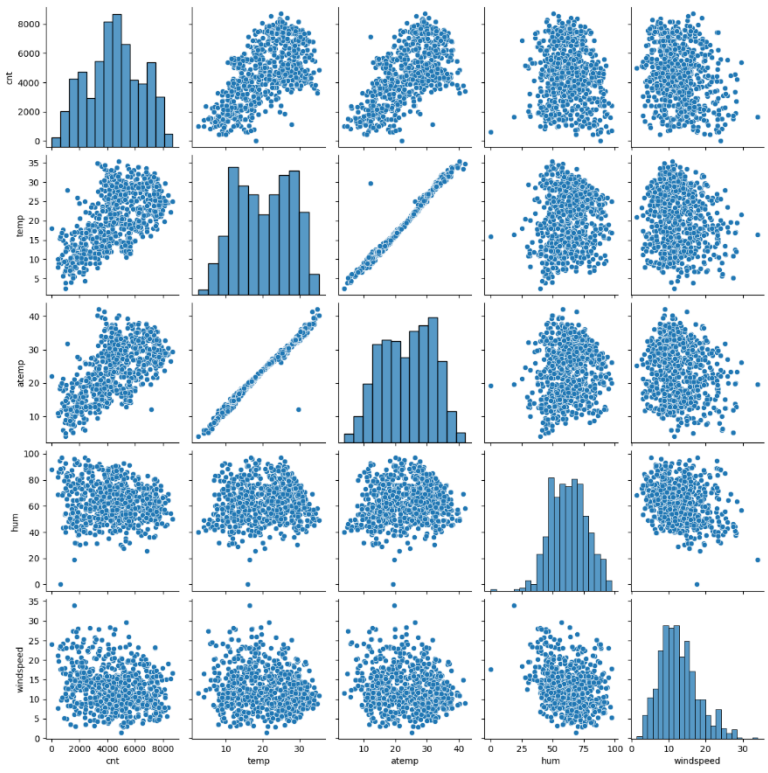
Answer :

Here in our assignment the target variable is ‘cnt’

And the rest numerical variable are 'temp', 'atemp', 'hum','windspeed'

After observing pair-plot; the variable ‘temp; has the highest correlation with the target variable ‘cnt’

Below is a pair plot



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

Answer:

I have Validated the assumptions of linear regression by checking the VIF, error distribution of residuals and linear relationship between the dependent variable and a feature variable.

Checking the VIF:

During model building we have checked each model's P-value and VIF:

- if the p-values >0.05 and the variable is insignificant we removed that variable ;
- along with checking the p-values we have checked for multicollinearity simultaneously for each model:

  for this we calculate VIF (Variance inflation factor) for each independent variable. VIF assesses the degree of multicollinearity among the independent variables. High VIF values indicate high correlation between predictors. If VIF is too high (usually above 10), it suggests that multicollinearity might be a concern.

  Addressing multicollinearity involve identifying the variable significance and then removing it according from the model.

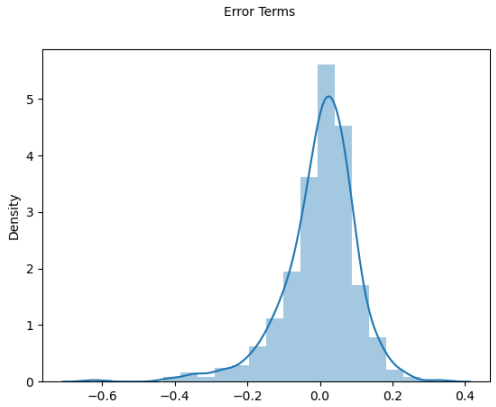
  In our case though temp has high VIF we kept it as it is important variable for inference. we removed

  Other variable with high VIF one at a time

Error distribution of residuals:

Residuals are the differences between the actual and predicted values. Checking the distribution helps ensure that the model's errors meet the assumptions of normality.

We ploted a histogram to check if the residuals follow a roughly normal distribution.

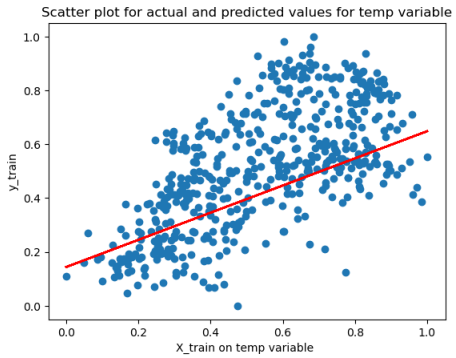


so we can see that the Error term distribution is centred around zero and approximately a normal distribution,so our assumption is valid

Linear relationship between the dependent variable and a feature variable:

It Checks if the relationship between what you're predicting and what you're using to predict is a straight line.

We plot a graph(for top Variable 'temp') that shows how the predicted values are relate to the real values. It should look like a line and it is looking like a line.



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

As per our final Model, the top 3 predictor variables that influences the bike booking are:

- **Temperature (temp)** - A coefficient value of ‘‘0.5043’ indicated that a unit increase in temp variable increases the bike hire numbers by ‘0.5043 units.
- **windspeed**: A coefficient value of ‘-0.1794’ is a negative values; indicated that, a unit increase in windspeed variable decreases the bike hire numbers by 0.1794 units.
- **Year (yr)** - A coefficient value of ‘0.2394’ indicated that a unit increase in yr variable increases the bike hire numbers by 0.2394 units.

So, it's suggested to consider these variables utmost importance while planning, to achive maximum Booking

The next best features that can also be considered are

- **season_winter:** A coefficient value of ‘0.0734’ indicated that, a unit increase in winter variable increases the bike hire numbers by 0.0734 units. People prefer to wait until the temperature are not that very high or extremely cold.
- **weathersit_moderate(weathersit 2):** A coefficient value of ‘-0.0680’ indicated that, a unit increase in variable, decreases the bike hire numbers by 0.0680 units.
- **mnth_sep:** A coefficient value of ‘0.0648’ indicated that, a unit increase in mnth_sep variable increases the bike hire numbers by 0.0648 units. Mostly climate in September with plenty of warm sunshine and clear blue skies for most of the month. September is a great month to have road trips as they balance the weather between warm and cold.

General Subjective Questions

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

Answer:

Linear regression is a supervised ML algorithm used for predicting the dependent variable based on one or more independent variables also called as predictor variables. The goal is to find the linear relationship that best describes the data.

The LR algorithm in detail:

Basics of Linear Regression:

1. **Model Representation:**
    - Linear regression assumes a linear relationship between the input variables (features) and the output variable (target).
    - The equation for a simple linear regression with one predictor variable is:

      $$y=mx+b,$$

      where

      y is the dependent variable, (plotted on y-axis)

      X is the independent variable, (plotted on x -axis)

      m is the slope,

      b is the intercept.

2. **Objective:**
    - The goal is to find the values of m and b that minimize the difference between predicted and actual values.
    - This is often done by minimizing the sum of squared differences (least squares method).
3. **Cost Function:**
    - The cost function measures the difference between predicted and actual values.
    - In linear regression, the common cost functions are mean squared error (MSE) or mean absolute error (MAE).
4. **Optimization Algorithm(Gradient Decent):**
    - An optimization algorithm, such as gradient descent, iteratively adjusts model parameters (*m* and *b*) to minimize the cost function.

Types of Linear Regression:

Linear regression models can be classified into two types depending upon the number of independent variables:

1. **Simple Linear Regression:**

    - Involves one independent variable.
    - Equation of Simple Linear Regression:

      $y = mx + b$

      OR

      $y = b0 + b1x$

2. **Multiple Linear Regression:**

    - Involves more than one independent variable.
    - Equation:
    - $y = b0 + b1x1 + b2x2 + \cdots \ldots \ldots \ldots \ldots bnxn$

      Where,

      y is dependent variable

      x1, *x*2, ...,*xn* are the independent variables.

      *b*0 is the intercept,

      b1,*b*2,...,*bn* are the coefficients

Steps for implementing Linear Regression:

1. **Data Collection and understanding the data**
    i) Importing required libraries for data analysis, manipulation and for data visualization like numpy ,pandas , seaborn & matplotlib
    ii) Preparing and refining the data to meet the standards required for exploratory data analysis. This involves addressing null values, ensuring appropriate formats, adjusting data types if necessary, and eliminating unnecessary rows or columns. It is essential to thoroughly clean the raw data before proceeding with visualization to ensure accurate insights.
2. **Data visualisation / EDA:**
    i) Visualizing numerical variables using scatter or pairplots in order to interpret business /domain inferences.
    ii) Visualizing categorical variables using barplots or boxplots in order to interpret business/domain inferences.
3. **Data preparation**:
    i) Converting categorical variables with varying degrees of levels into numerical(binary mostly) dummy variables so that these variables can be used during model building in order to contribute to the best fitted line for the purpose of better prediction.
    ii) Splitting the data into training and test sets

iii) Splitting the data into two sections train set and test set. Generally, the train-test split ratio is 70:30 or 80:20.
iv) Rescaling the trained model to normalize the range of numerical variables with varying degrees of magnitude.

4. **Model Building:**
   Building a linear model
   i) Forward Selection: We start with null model and add variables one by one. These variables are selection on the basis of high correlation with target variable.
   ii) Backward Selection: We add all the variables at once and then eliminate variables based on high multicollinearity (VIF>5) or insignificance (high p- values generally p-values>0.05).
   iii) RFE or Recursive Feature Elimination is more like an automated version of feature selection technique where we select that we need "m" variables out of "n" variables and then machine provides a list of features with importance level given in terms of rankings. A rank 1 means that feature is important, rank 2 means next important features after rank 1 and so on.
   Generally we select 10 to 15 features variables
5. **Residual analysis of the train data:**
   It tells us how much the errors (y_actual — y_pred) are distributed across the model.
   A good residual analysis will signify that the mean is centred around 0.and Error terms are approximately normally distributed
6. **Making predictions using the final model and evaluation:**
   i) We will predict the test dataset by transforming it onto the trained dataset
   ii) Divide the test sets into X_test and y_test and calculate r2_score of test set. The train and test set should have similar r2_score.
   A difference of 2–3% between r2_score of train and test score is acceptable as per the standards.

Assumptions of Linear Regression:

1. **Linearity:**
   - Assumes a linear relationship between predictors(x) and the target(y).
2. **Normality Distribution of Error terms:**
   - Assumes that residuals (errors) are normally distributed.
3. **Independence:**
   - Assumes that observations are independent of each other.
4. **Homoscedasticity:**
   - Assumes constant variance of errors across all levels of predictors.

Advantages and Disadvantages:

Advantages:

- Simple to understand and implement.
- Interpretable coefficients provide insights into variable relationships.

Disadvantages:

- Assumes linearity, which may not always hold.
- Sensitive to outliers.
- Can struggle with complex relationships.

In summary, linear regression is a foundational and widely used algorithm for predicting continuous outcomes. Its simplicity and interpretability make it a valuable tool in various domains, but it is essential to be aware of its assumptions and limitations.

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

Answer :

Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyze it and build your model.

Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but have very different distributions so they look completely different from one another when you visualize the data on scatter plots.

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

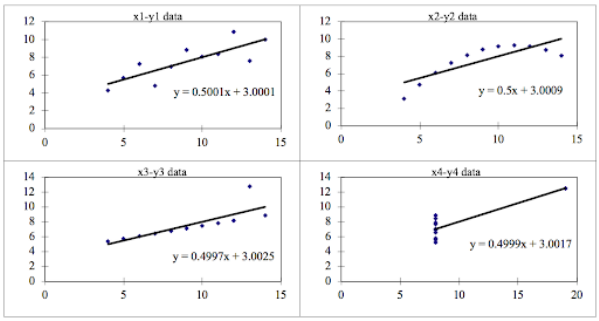
We can define these four plots as follows:

Let is consider below four data model with their summary statistics:

Anscombe's Data									
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

Here we can see all the four models have equal mean, SD and r values.

But, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as you can see below:



Refering to above graphs we can describe the above four data sets as:

Data Set 1(x1,y1): It fits the linear regression model pretty well.

Data Set (x2,y2): The data points are not linear and so it cannot fit the linear regression model

Data Set 3: It shows the outliers present in the data set, which cannot be handled by the linear regression model.

Data Set 4: In shows the outliers present in the data set, which also cannot be handled by the linear regression model.

As you can see, Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

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

Answer:

Pearson's correlation coefficient, often represented as Pearson's R, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It assesses how well the relationship between the variables can be described by a straight line.

The coefficient ranges from -1 to 1:

Positive R: Indicates a positive correlation, meaning as one variable increases, the other tends to increase as well. A value close to 1 signifies a strong positive correlation.

Negative R: Indicates a negative correlation, suggesting that as one variable increases, the other tends to decrease. A value close to -1 indicates a strong negative correlation.

R close to 0: Suggests a weak or no linear correlation between the variables.

This measure is commonly used in various fields to assess connections between variables, but it assumes that the relationship is linear and is influenced by outliers.

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

Answer :

Scaling/feature scaling is a process in data preparation where all numerical features/variable of a dataset are transformed to a standard range or distribution.

When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret.

So we need to scale features because of two reasons:

1. Ease of interpretation

2. Faster convergence for gradient descent methods

There are two very popular method are used for scaling the variables:

Normalised /MinMax Scaling:

The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

The formula for Normalised scaling is :

$$X_{\text{normalized}} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

Normalised scaling is Used when the independent variables have varying ranges and need to be constrained to a specific interval.

Standardize scaling:

The variables are scaled in such a way that their mean is 0 and standard deviation is 1.

The formula for Standardizing scaling is :

$$X_{\text{standardized}} = \frac{X - \text{mean}(X)}{\text{std}(X)}$$

Standardize scaling is used when independent variables have different units and scales, and we want them to be comparable.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F statistic, p-values, R-square, etc.

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

Answer :

The value of VIF is infinite when there is a perfect correlation between the two independent variables of the dataset. So in this case the R-squared value is 1 .

Now we know the equation of VIF

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

So when R^2 is 1 , VIF becomes infinity.

Infinite VIF values can cause numerical instability in regression models.

It indicates that the correlation between certain variables is so high that the model cannot distinguish their individual effects.

This suggest that is there is a problem of multi-collinearity and one of these variables need to be dropped in order to define a working model for regression.

6. 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, short for quantile-quantile plot, is a graphical tool used to assess if a dataset follows a particular theoretical distribution. It is commonly used in statistics to check the normality assumption of a dataset by comparing its quantiles to the quantiles of a theoretical normal distribution.

In the context of linear regression, a Q-Q plot can be employed to assess whether the residuals of a regression model are normally distributed.

How to Understand a Q-Q Plot:

Perfect Line:

If the points on the graph make a nice straight line, it means our residuals are behaving like they should in a normal world.

Curves or Turns:

If the points deviate from the straight line, it suggests our residuals might not be perfectly normal.

In simple language A straight line means our model is doing well. Curves or turns suggest we might need to fix something.

How to Use a Q-Q Plot in Simple Steps:

Calculate Residuals:

Find the differences between our predictions and the real outcomes.

Draw the Q-Q Plot:

Make a graph comparing our residuals to what they should look like in a perfect world.

Check the Line:

If the points form a nice line, things are good. If not, we might need to look closer.

What It Tells Us:

A straight line means our model is doing well. Curves or turns suggest we might need to fix something.

Importance of Q-Q plot in Linear Regression:

The Q-Q plot helps us see if the assumption made by linear regression model that residuals are normally distributed,is holds valid.

If the data points don't make a straight line, it tells us there might be a problem with our model or data.

Limitation of Q-Q plot :

For small sets of data, Q-Q plots might not be as helpful.

We should Combine Q-Q plots with other checks to get a clearer picture of how well our model is working.

In a nutshell, a Q-Q plot is a valuable tool in linear regression, helping us make sure our model is doing its job correctly and spotting any areas that might need fixing.