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 -
   The analysis of categorial variables from the dataset is done using box-plot and the inferences about their effect on the dependent variable (‘cnt’) are as follows:

   - **Season -** Fall and Summer season shows more count of bikes rented.
   - **Month (mnth) -** The demand for shared bikes started increasing from the month of March, demands saw its peak in months of (June to October) and then demand started decreasing in months of November and December.
   - **Weathersit -** Users have more booking on Clear weathers, and no record of booking was found on Heavy weathers (Heavy Rain, Thunderstorm).
   - **Holiday -** In holidays the demand is less as compare to no holidays. Maybe people are at home on holidays and not going out.
   - **Year (yr) -** The demand for shared bike have increased in 2019 as compare to 2018, if the trend continues upcoming years can be profitable for the company
   - **Weekday -** Weekday shows similar results for all days
   - **Workingday** - Working or Non-Working days shows similar behaviour.

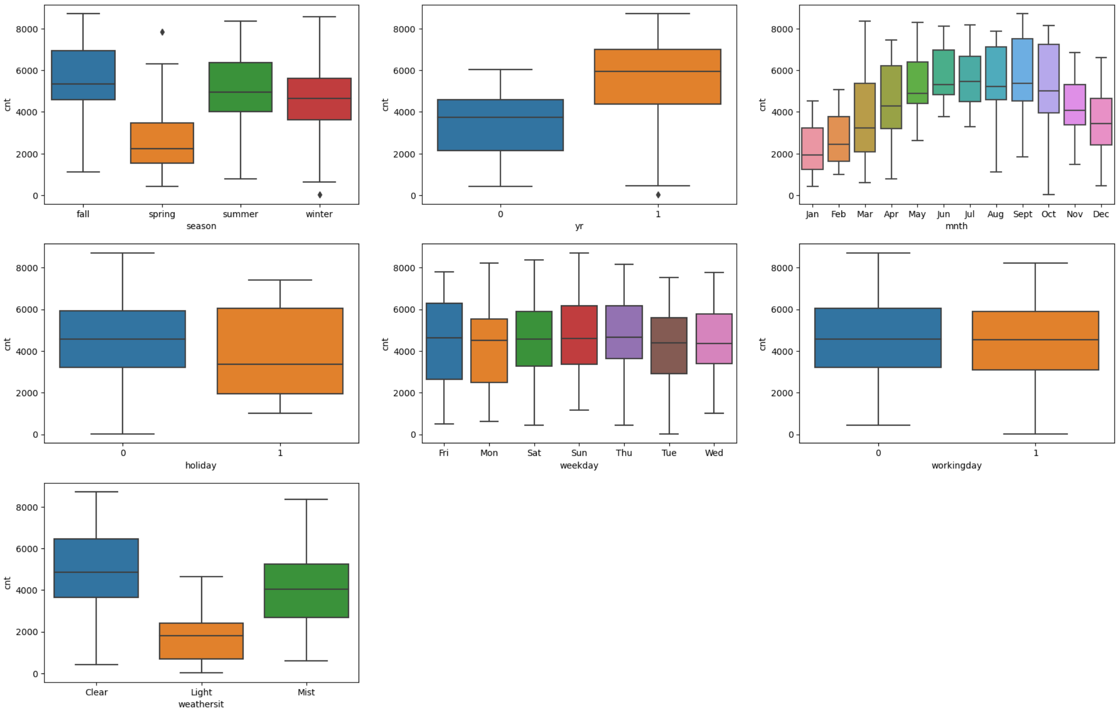


Figure 1 : Analysis of Categorical Variable

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

Answer -

Syntax - *drop_first : bool, default False*
Whether to get k-1 dummies out of k categorical levels by removing the first level.
While creating dummy variable if we don't use **drop_first=True**, there is one extra dummy variable is created, which can result in **multicollinearity** of dummy variables with each other, as well as **difficulties in interpretation**. So if we can drop the first dummy variable, we can easily avoid this problem.
For example :
Let's say we have a categorical column named 'Relationship_Status' which contains 3 levels ('Single','In a Relationship' and 'Married'). We can create a dummy table like this :

If we use, **drop_first=False,** we get the below results :
(This method will result in **multicollinearity** of dummy variables with each other)

Relationship_Status	Single	In a Relationship	Married
Single	1	0	0
In a Relationship	0	1	0
Married	0	0	1

If we use, **drop_first=True,** we get the below results :
(This way we can reduce redundancy and Multicollinearity in the model)

Relationship_Status	In a Relationship	Married
Single	0	0
In a Relationship	1	0
Married	0	1

As we can see that there is no need to define three different variables (in columns, Single is missing, by setting **drop_first=True** in dummy creation). If we drop a level, 'Single' we will still be able to explain all the three levels. If ('In a Relationship' and 'Married' are (0,0) it is obvious that the row is for 'Single').

---

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

   Answer -

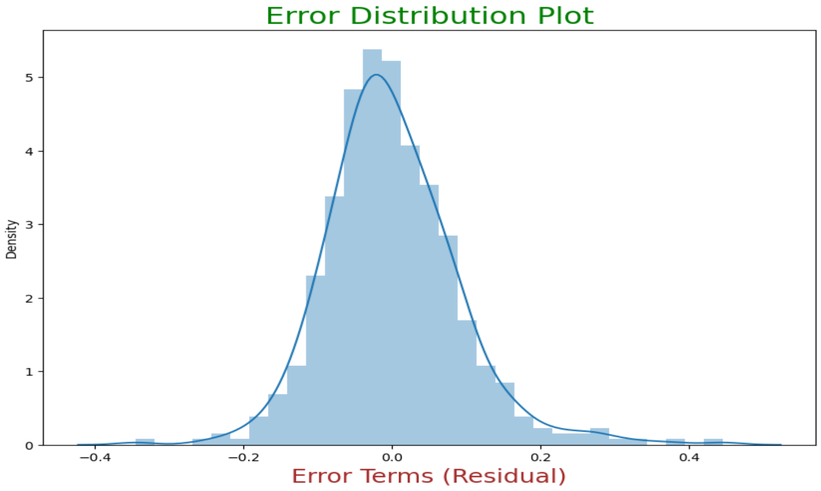
   ‘temp’ variable has the highest correlation with the target variable.

---

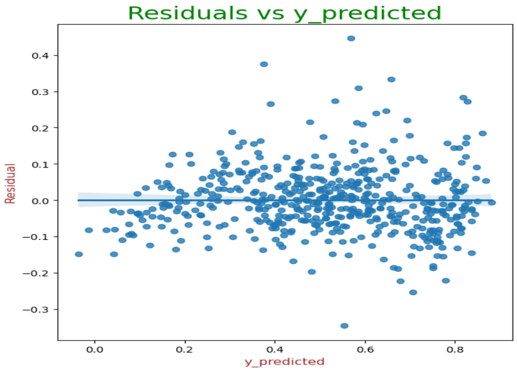
4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

   Answer -

   1. **Normal Distribution of error terms**
      The residuals are normally distributed around zero in our distribution plot



   2. **Independence of error terms**
      No patterns identified in the plot (Residuals VS y_predicted)



   3. **Multicollinearity**
      While VIFs calculation, high correlated variables were dropped (with VIF >10).
   4. **Homoscedasticity**
      The variance of the residual, or error term, is constant after plotting regplot.

---

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

   Answer -
   Top 3 features contributing significantly are :
   1. **'temp'-** (1 unit increase in 'temp' will increase 'cnt' (count of bikes) by 0.4387)
   2. **'yr' -** (1 unit increase in 'yr' will increase 'cnt' (count of bikes) by 0.2345)
   3. **'weathersit_Light' -** (1 unit increase in 'weathersit_Light' will increase 'cnt' (count of bikes) by - 0.2917).

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**
   Answer -
   A linear regression algorithm is a supervised learning method which attempts to explain a linear relation between a Target variable (dependant variable) which is continuous in nature and one or more Predictor variables (independent variable) using a straight line. In regression, the best fit line is a line that fits a given scatter plot (between Target variable on y-axis and Independent variable on x-axis) in the best way which tries to minimize the error.

   There are two types of Linear Regression :
   - **Simple Linear Regression (SLR)**
     This is used when the number of independent variables is 1 (X = 1 variable)
   - **Multiple Linear Regression (MLR)**
     This is used when the number of independent variables is more than 1 (X > 1 variable). This is similar to SLR but with multiple independent variables.

   Mathematically, the Equation of Linear Regression is given as :
   $yi = \beta o + \beta 1X1 + \beta 2X2 + \beta 3X3 + ...\beta nXn + \mathcal{E}$

   Where,
   yi = Target variable, i = n observations
   βo = y-intercept (constant)
   $\beta 1$ = Slope coefficient for 1st feature
   βn = Slope coefficient for nth feature
   Xn **=** Independent variables
   $\mathcal{E}$ **=** Residual (error terms)

   Example :

   Let's say a car salesman wants to predict the price of old used cars on the basis of it age (years used), physical condition (how well maintained) and mileage. So he decides to collect past data of a number of used cars on these variables.

He can use that data to fit a MLR model where price (continuous variable) of the car is the dependent variable, and the age, physical condition & mileage are the independent variable. Let's assume that the model has come up with an equation after making predictions about the price of the car using the model. We can use that equation to finally find the price of the used car. For example, if he has to sell a used car that is 5 years old, has 800,000 miles on it, and is in good condition, then he can use that equation to find the best price for that car.

Assumptions of Linear Regression :

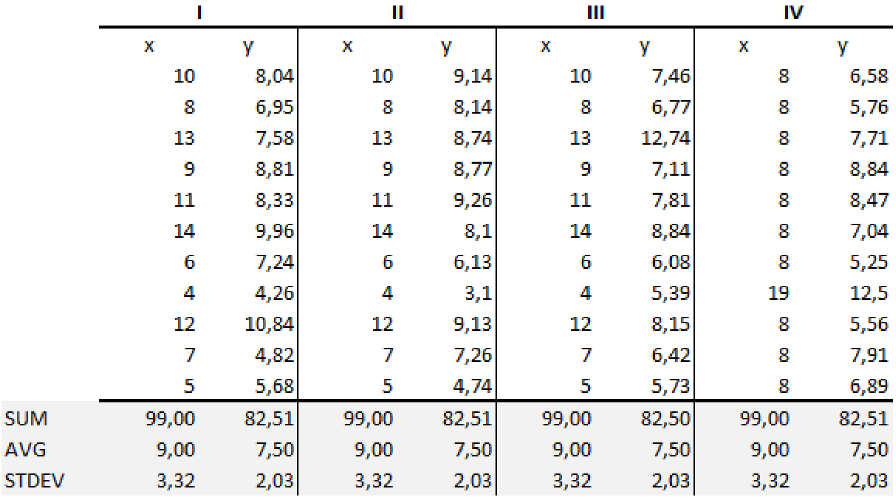
- Linear Relationship between target variable and predictor variables.
- Error terms are independent of each other.
- Error terms are normally distributed.
- Homoscedasticity (Error terms have constant variance).

---

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

Answer -

The Anscombe's quartet was created by statistician Francis Anscombe. It contains four datasets, each dataset contains 11 (x, y) pairs. Each of which has the same summary statistics, but when they are graphed, their distribution looks very different from each other. Sometimes, even the statistical information summarized from the data may mislead you to wrong conclusions. Better plot the graphs to understand the data if there are any patterns to it.

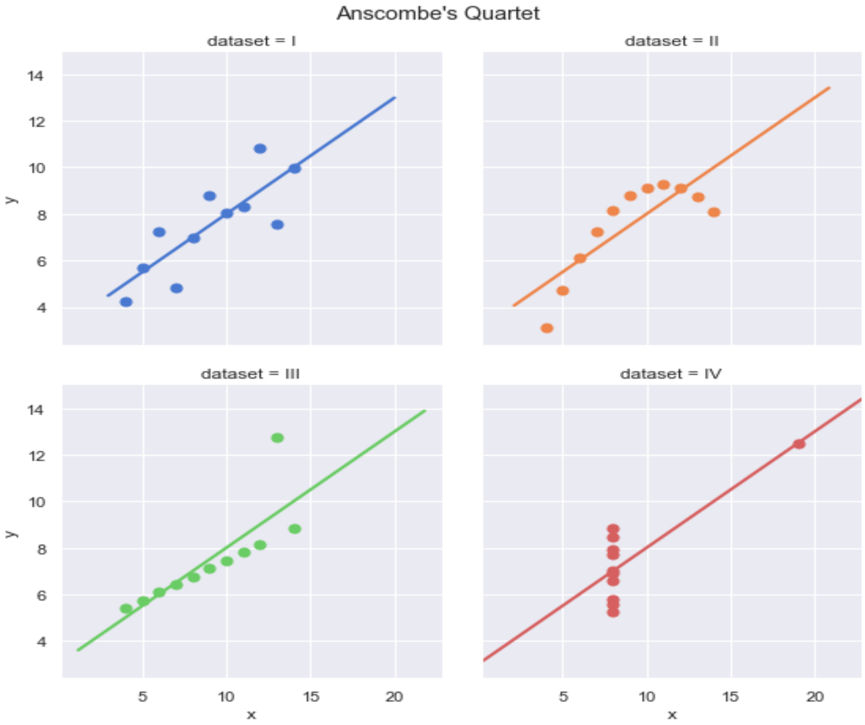


	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	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
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

As we can see, the summary statistics in above data

- The mean x & y is (9, 7.5) respectively in all the 4 datasets.
- The Standard deviation of x is 3.32 and for y is 2.03 in all the 4 datasets.
- Also, the correlation coefficient between x & y is 0.816 for each dataset.

When we plot these datasets we see a different story even if their statistics summary and regression line is the same. The plots are below as follows:



- **Data-set I** - consists of a set of (x, y) points that represent a linear relationship with some variance.
- **Data-set II** - shows a curve shape but doesn’t show a linear relationship.
- **Data-set III** - looks like a tight linear relationship between x and y, except for one large outlier.
- **Data-set IV** - looks like the value of x remains constant, except for one outlier.

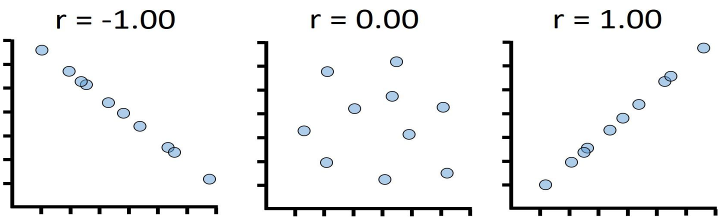
---

3. What is Pearson’s R? (3 marks)

Answer -

Pearson's R, also known as Pearson's correlation coefficient, is a measure of the linear association between two variables. It, ranges from -1 to 1,measures the strength and direction of the relationship between two variables.
The coefficients tell us as follows:

- r = -1 indicates a **perfect negative** linear relationship, the slope is negative.
- r = 0 indicates **no linear relationship**.
- r = 1 indicates a **perfect positive** linear relationship, the slope is positive.



When **r > 0 (positive)**, it means that when the value of one variable increases (on x-axis), the value of other variable on y-axis will also increase.
When **r < 0 (negative)**, it means that when the value of one variable increases (on x-axis), the value of other variable on y-axis will decrease.

Pearson correlation coefficient formula is given as:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
$\mathbf{y_i}$ = values of the y-variable in a sample
$\mathbf{\bar{y}}$ = mean of the values of the y-variable

---

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

   Answer -

   - **Scaling** is the process done in data preparation stage in model building which transforms the input features of ML model so that they have a common comparable scale and the data points are generalized. This also speeds up the calculations in an algorithm.
   - The reason for doing this is to ensure that the model is not affected by the large value ranges of certain features. For example: The dataset mostly contains variables highly varying in magnitudes, units. If we don't perform scaling, then the algorithm only takes magnitude into account and not units, hence can lead to incorrect modelling. (Let's say a dataset contains price of House (in dollars) and number of bedrooms (1,2,3) the magnitude and the units are totally different so to have a common comparable scale/magnitudes we perform scaling). Scaling doesn't impact the models parameters like **p-values, R-squared, etc**, it just affects the coefficients.

   <u>Difference between normalized scaling and standardized scaling:</u>

Normalization : It brings all the data in the range of 0 and 1. Normalization is useful when the distribution of the variable is not known. It is affected by outlier.

- **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Standardization : It replaces the values by their Z scores. It brings all the data into a standard normal distribution which has **mean (μ) zero** and **standard deviation one** (σ). It is less affected by outliers as compare to normalization.

- **sklearn.preprocessing.scale** helps to implement standardization in python.

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

---

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

Answer -

VIF formula is given by : $VIF_i = \frac{1}{1 - R_i^2}$, Where R^2_i is the R-squared,

- VIF (Variance Inflation Factor) is a measure of how much the variance of an estimated regression coefficient is increased due to collinearity (correlation) among the independent variables. It detects multicollinearity in regression analysis.
- **VIF = 1** represents that there is **no correlation** among the independent variables, while a **VIF > 1** represents that **there is correlation**. The VIF value can vary from 1 to infinity, with higher values representing a higher degree of collinearity. The common heuristic for VIF is that, VIF greater than 10 is considered high.
- Mathematically, **VIF is infinity** when **R-Squared equal to 1,** which means a perfect correlation between two independent variables. In this case, the estimated regression coefficients become unreliable and infinitely large. This can be solved if we drop one of the independent variable from the model, **as it can be linearly represented by the other variable.**

---

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

Answer -

Using Q-Q plots, one can determine if two data sets are derived from the same distribution. A Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. The 1st quantile is that of the variable we are testing the hypothesis for, & the 2nd is the actual distribution we are testing it against.

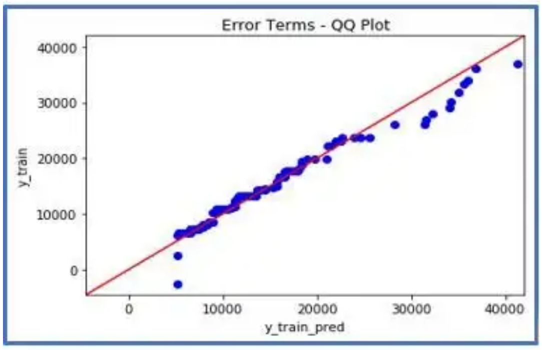
➢ For example, if we are testing if the distribution of Height of employees in our team is normally distributed, we are comparing the quantiles of our team members' Height vs quantile from a normally distributed curve. If 2 quantiles are sampled from the same distribution, they should roughly fall on a straight line.

In Linear Regression, this plot has **importance** in helping when we have training and test dataset separately, and we can use Q-Q plot to confirm that both the data sets are from populations with same distribution.

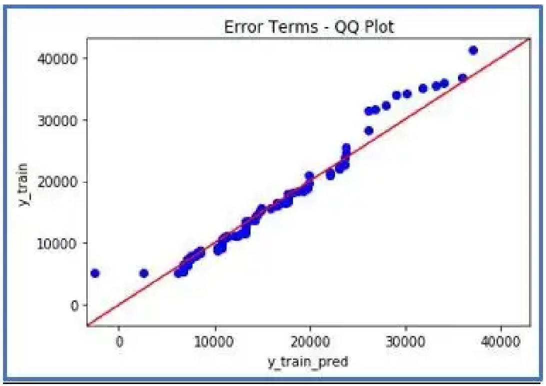
Below are the possible **interpretations** for two data sets.

1. **Similar distribution:** If all point of quantiles lies on or close to straight line at an angle of 45 degree from x-axis.

2. **Y-values < X-values:** If y-quantiles are lower than the x-quantiles. The graph can be seen below.



3. **X-values < Y-values:** If x-quantiles are lower than the y-quantiles.



4. **Different distribution:** If all point of quantiles lies away from the straight line at an angle of 45 degree from x-axis.

Uses of Q-Q plot :

1. To check if 2 data sets have common location and scale.
2. It can be used with sample sizes also.
3. The presence of outliers can all be detected from this plot.
