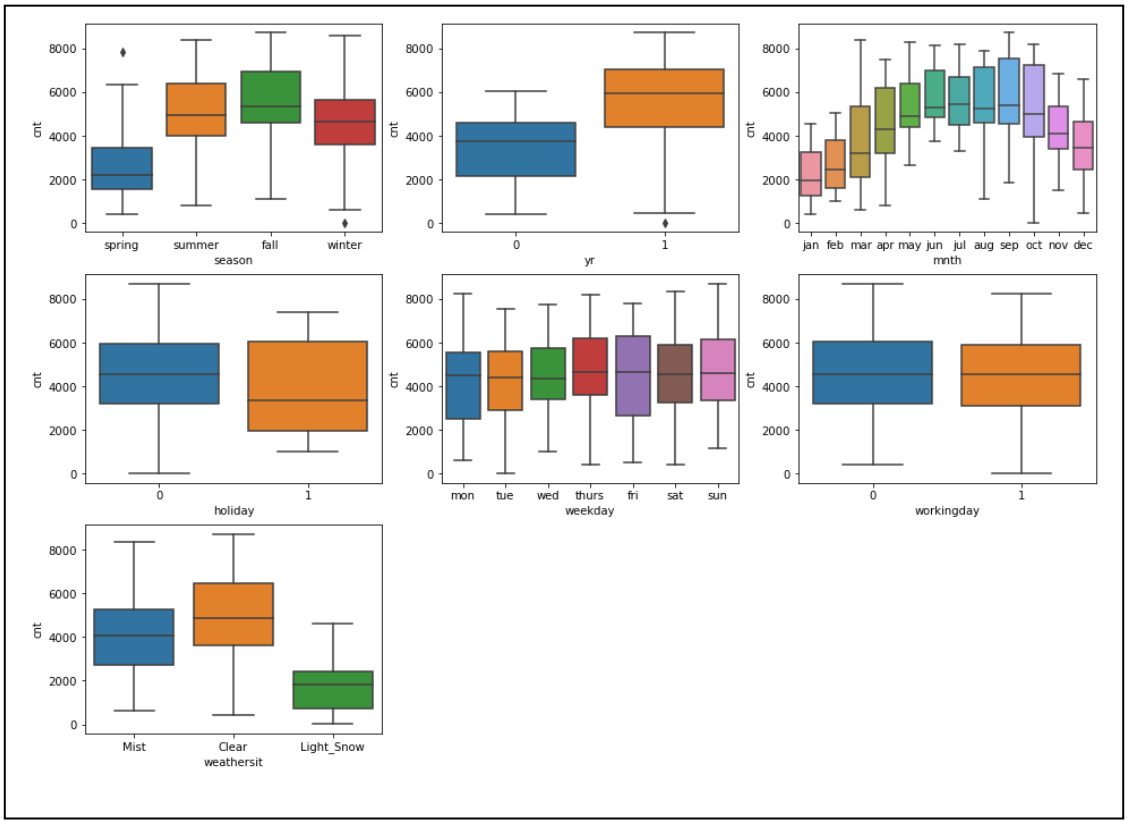
Submitted By : Sushant Sarswat (DS C-38)

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)**

Sol: The inferences drawn from my analyses of the effect of categorical variables on dependent variables are as follows:-



(a) Fall has the highest bookings amongst all seasons for the period of two years followed by summer and winter. This indicates season to be good predictor for demand of bikes.

(b) The demand for bikes in Year 2019 is more than for Year 2018. This indicates an upward trend with passing years.

(c) The demand for bikes rises steadily from Jan till Jun, is stable till Sept and gradually declines till Dec thereafter. This is in consonance with the observations of season.

(d) The demand for bikes is observed to be lower on holidays. Therefore, impetus of bike sharing business should be on working days.

(e) Overall median across all days is the same. Also, the median is same for working and non-working days.

(f) The demand of the bikes is highest on the days when weather is clear, followed by misty days and least on light snow days. There seems to be no demand on days with heavy rain.

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

Sol: A variable consisting of 'n' levels can be represented by 'n-1' dummy variables. Therefore, even if we remove the first column, we can still represent the entire data. This is especially important for optimisation, since total number of combinations of features are given by 2^n and can be quite substantial for models with higher number of features.

e.g. For a variable 'Wealth Status', we can have a dummy table as follows:-

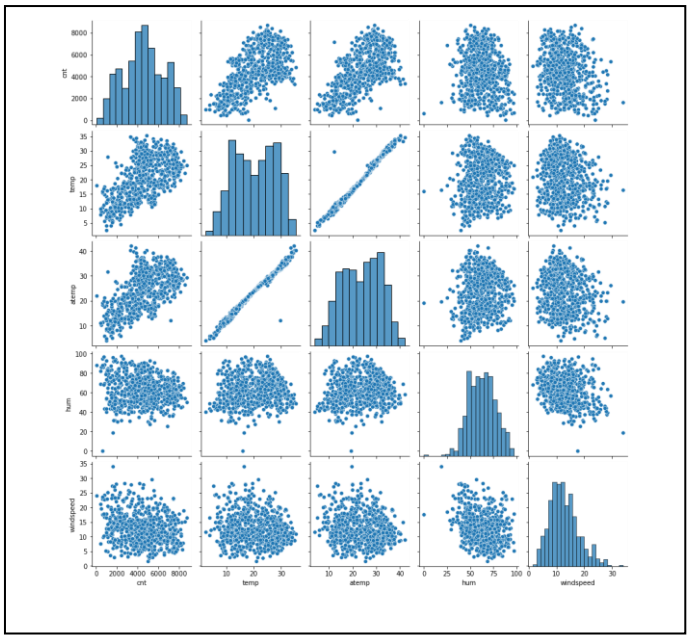
Wealth Status	Rich	Average	Poor
Rich	1	0	0
Average	0	1	0
Poor	0	0	1

The above information can still be represented without any loss if we drop the first column as follows:-

Wealth Status	Average	Poor
Rich	0	0
Average	1	0
Poor	0	1

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

Sol:

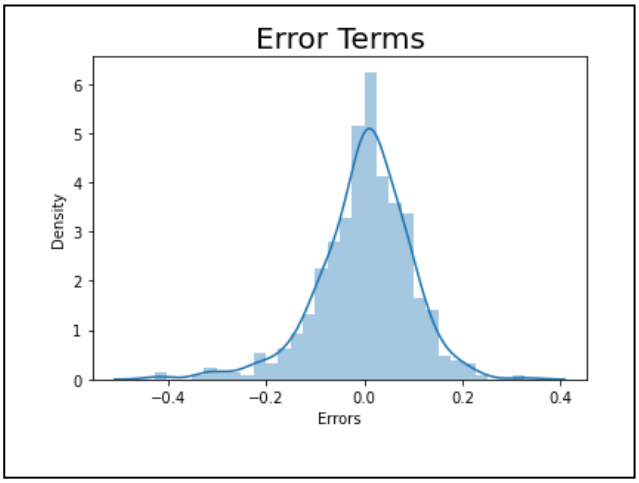


The highest correlation with target variable by looking among numerical variables in pair plot is of those with 'temp' and 'atemp', each with a value of 0.63.

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

Sol: I validated the assumptions of Linear regression after building the model on training set as follows:-

(a) To validate the assumption '**Error terms are normally distributed with mean zero (Not X and Y)**', I undertook the Residual Analysis of Training Data, wherein the histogram revealed that the Errors are normally distributed with mean of errors centered at 0.



(b) The linear relationship between 'cnt' (target variable) and independent numerical variables was asserted with the help of pair plot, as indicated in Sol 3 ibid.

(c) VIF of the model was used to ascertain the assumption that '**There is no Multicollinearity between predictor variables**'. The VIF values of features was observed to be less than 5.0.

	Features	VIF
2	windspeed	4.72
1	temp	4.09
0	yr	2.03
3	season_spring	1.82
8	weathersit_Mist	1.50
4	season_winter	1.36
5	mnth_mar	1.20
6	mnth_sep	1.15
7	weathersit_Light_Snow	1.07

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

Sol: The following three features are contributing significantly towards explaining the demand of shared bikes:-

(a) temp (0.4233)

(b) yr (0.2320)

(c) weathersit_Light_Snow (-0.3136)

General Subjective Questions

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

Sol: A linear regression algorithm tries to establish a relationship between dependent and independent variables with the help of a straight line and is applicable only to numerical variables. The following are broadly the steps undertaken whilst performing the linear regression:-

(a) The dataset is checked for its correctness.

(b) The dataset is divided into training and test data.

(c) The Training data is divided into Features(independent variables) and target (dependent variables).

(d) Internal functions available in Python libraries statsmodels and sklearn are used to arrive at a linear model fitted using the training data. The algorithm of Gradient Descent are used in the backend to find the coefficients of the line with best fit. This algorithm works on the principle of minimising the cost function (e.g. Residual sum of squares).

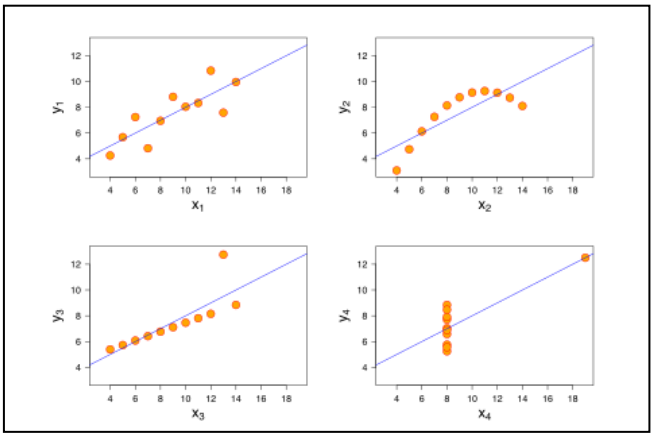
(e) In case of multiple features, the predicted variable is a hyperplane instead of a line and therefore , the model is depicted as follows:-

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots\ldots\ldots\ldots + \beta_n X_n$$

(f) The performance of the model is evaluated with the help of test data and assumptions are ascertained for their correctness.

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

Sol: Anscombe's quartet comprises of four datasets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. The simple statistics comprises of sample variance of X and y, mean, correlation coefficient, linear regression line and R-square value. These quartets exhibit that multiple datasets with many similar statistical properties can still be vastly different from each other when graphed.



Source: https://en.wikipedia.org/wiki/Anscombe%27s_quartet

(a) Top LHS : Simple Linear relationship

(b) Top RHS : Non-linear relationship, Not distributed normally, Correlation coefficient irrelevant.

(c) Bottom LHS : Linear but different regression line because of presence of outliers.

(d) Bottom RHS : No linear relationship. Statistics adjusted due to outliers.

Anscombe's quartet are used to emphasize upon the importance of visualisation and removal of outliers prior to data analysis.

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

Sol: Pearson's R is used to measure the strength of association of two variables. Mathematically, it is achieved by dividing the covariance of two variables by the product of their standard deviation. It has a range from -1 to +1.

(a) A value of 1 indicates a positive linear correlation. (i.e. if one variable increases, the other also increases)

(b) A value of 0 indicates no linear correlation. However, non-linear correlation may exist.

(c) A value of -1 indicates negative correlation. (i.e. if one variable increases, the other decreases)

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

Sol: Scaling of a variable means a method to keep the coefficients within a certain range. It is a pre-processing step in linear regression. The major advantage of scaling is that it makes the process of Gradient Descent faster as well as for easy comprehension to understand the effect of each feature of the model on the target variable.

Factor	**Normalised Scaling (MinMax Scaling)**	**Standardised Scaling**
Formula	X_new = (X - X_min)/(X_max - X_min)	X_new = (X - mean)/Std

Factor	Normalised Scaling (MinMax Scaling)	Standardised Scaling
Factors used for Scaling	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
Terms of Usage	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
Outlier Impact	It is really affected by outliers	Does not get affected by outliers
Distribution	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
Python Libraries	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

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

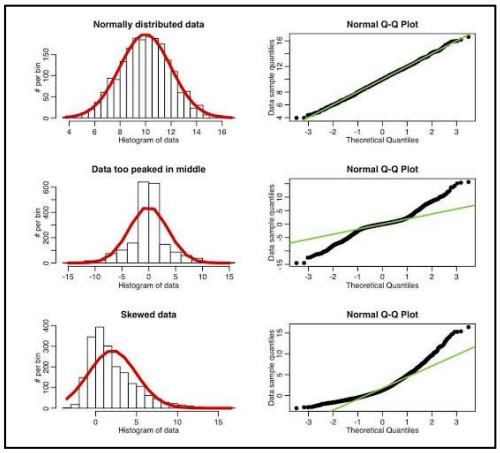
Sol: Formula of VIF is

$$\mathbf{VIF_i = 1 / (1-R_i^2)}$$

For VIF_i to be infinite, denominator has to be 0 which will happen when $R_i^2 = 1$. This is possible when there is perfect correlation between features.

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

Sol: A Q-Q plot is a scatter plot of two sets of quantiles against each other. It is used to ascertain if these two sets of data came from the same distribution. It is a visual check, wherein if the data is from the same source, the plot will appear as a line. Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution, Uniform Distribution, Exponential Distribution or even Pareto Distribution, etc.



Source:
https://towardsdatascience.com/q-q-plots-explained-5aa8495426c0