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)

Ans: The categorical variable in the dataset were season, weathersit, holiday, mnth, yr and weekday. These were visualized using boxplot .These variables had the following effect on our dependent variable:-

1. Season –

- The boxplot showed that spring season had least value of cnt whereas fall had maximum value of cnt. Summer and winter had intermediate value of cnt.
- Almost 32% of the bike booking were happening in fall with a median of over 5000 booking (for the period of 2 years). This was followed by summer & winter with 27% & 25% of total booking.
- This indicates, season can be a good predictor for the dependent variable.

2. **Weathersit** –

- There are no users when there is heavy rain/ snow indicating that this weather is extremely unfavorable. Highest count was seen when the weathersit was Clear, Partly Cloudy.
- Almost 67% of the bike booking were happening during ‘Clear-Partly cloudy’ with a median of close to 5000 booking (for the period of 2 years). This was followed by Mist-Cloudy with 30% of total booking.
- This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.

3. Holiday –

- Rentals reduced during holiday.
- Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is clearly biased.
- This indicates, holiday **CANNOT** be a good predictor for the dependent variable.

4.Mnth –

- September saw highest no of rentals while December saw least. This observation is on par with the observation made in weathersit. The weather situation in December is usually heavy snow.
- Almost 10% of the bike booking were happening in the months may,jun,jul,aug & sept with a median of over 4000 booking per month.
- This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.

5. Yr –

- The number of rentals in 2019 was more than 2018.
- Almost 99% of the bike booking were increased in year with median of close to previus year booking (for the period of 2 years).
- This indicates, yr can be a good predictor for the dependent variable.

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

Ans:

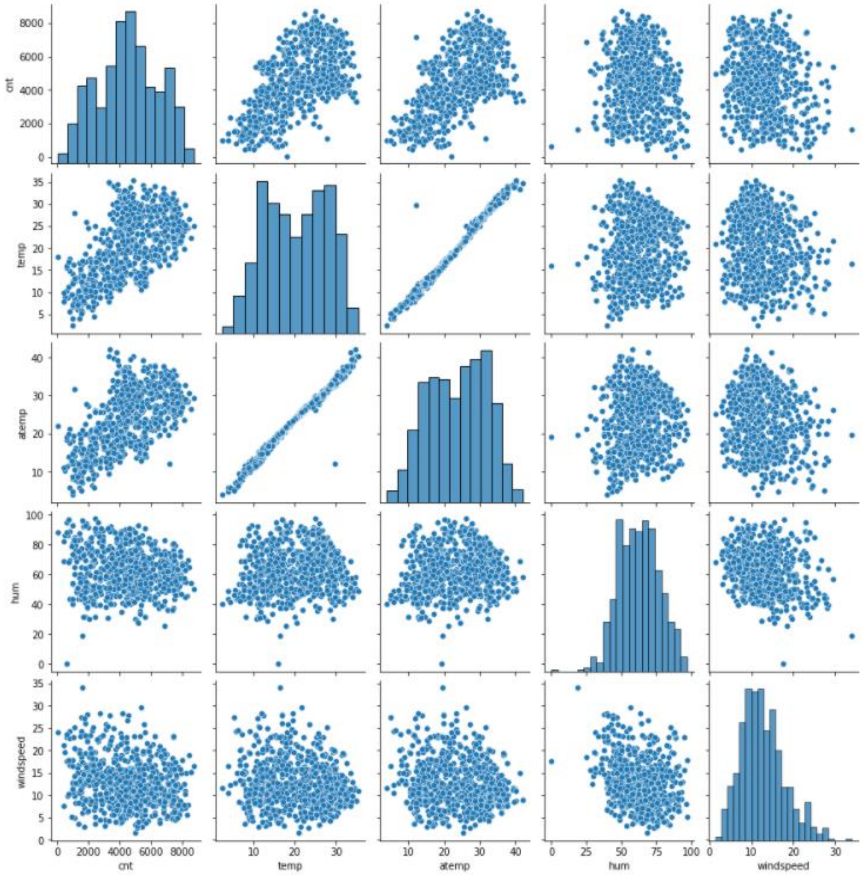
If you don't drop the first column then your dummy variables will be correlated (redundant). This may affect some models adversely and the effect is stronger when the cardinality is smaller.

For example:

- Iterative models may have trouble converging and lists of variable importance may be distorted.
- Another reason is, if we have all dummy variables it leads to Multi-collinearity between the dummy variables. To keep this under control, we lose one column.

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

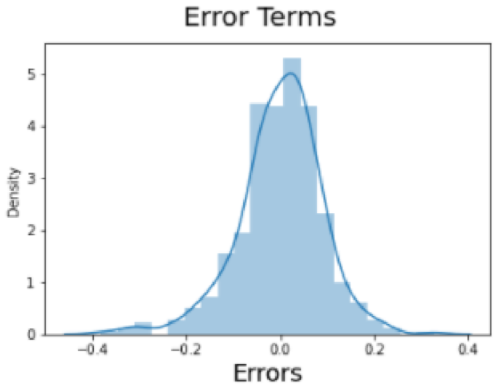
Ans:



"temp" and "atemp" are the two numerical variables which are highly correlated with the target variable (cnt).

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

Ans:



- Residuals distribution should follow normal distribution and centred around 0. (mean= 0). We validate this assumption about residuals by plotting a distplot of residuals and see if residuals are following normal distribution or not. The above diagram shows that the residuals are distributed about mean = 0
- Error terms seem to be approximately normally distributed, so the assumption on the linear modeling seems to be fulfilled.

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

Ans: The top 3 Features contributing significantly are:

I. temp with co-efficient **0.597749**

II. yr with co-efficient **0.227954**

III. weathersit_Light Snow & Rain with co-efficient **-0.231830**

General Subjective Questions

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

Ans :

- Linear regression is a predictive analytic technique, which is a relationship between one dependent variable and one or more independent variables.
- Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values. Linear Regression is the most basic form of regression analysis. Regression is the most commonly used predictive analysis model.

  Linear regression is based on the popular equation "y = mx + c".
- It assumes that there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x).
- In regression, we calculate the best fit line which describes the *relationship between the independent and dependent variable*.
- Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal/categorical etc.
- Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error.
- In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term.

Regression is broadly divided into simple linear regression and multiple linear regression.

- **Simple Linear Regression: SLR** is used when the dependent variable is predicted using only **one** independent variable.
- **Multiple Linear Regression: MLR i**s used when the dependent variable is predicted using multiple independent variables.

The equation for MLR will be: $y_i=\beta_0+\beta_1 x_{i1}+ \beta_2 x_{i2} +...+\beta_p x_{ip}+ \epsilon$

Where, for i= n observations:

yi = dependent variable (Output Variable)

xi = explanatory variables (Predictor Variable)

β0 = Intercept (the constant term)

β1 = coefficient of x1 variable.

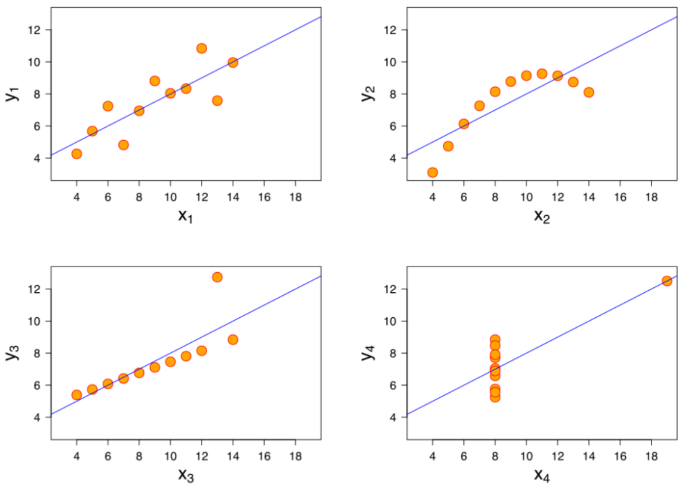
β2 = coefficient of x2 variable and so on…

ϵ = the model's error term (also known as the residuals)

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

Ans:

Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasize both the importance of graphing data before analyzing it and effect the of outliers and other influential observations on statistical properties.



- The *first scatter plot* (top left) appears to be a simple linear relationship.
- The *second scatter plot* (top right) is not distributed normally; while there is a relation between them, it's not linear.
- In the *third scatter plot* (bottom left), the distribution is linear, but should have a different regression line the calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the *fourth scatter plot* (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

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

Ans:

Pearson's r is a numerical summary of the strength of the linear association between the variables. It value ranges between **-1 to +1**. It shows the linear relationship between two sets of data. In simple terms, it tells us can we draw a line graph to represent the data?

○ r = 1 means the data is perfectly linear with a positive slope

○ r = -1 means the data is perfectly linear with a negative slope

○ r = 0 means there is no linear association

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

Ans:

Scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data preprocessing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.

In other words, it is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modeling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Normalized Scaling:

- Normalized Scaling also known as Mix/Max scaling
- It brings all of the data in the range of 0 and 1
- Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution.
  Formula: $\mathbf{X_{sc} = X - X_{min} / (X_{max} - X_{min})}$

Standardized Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has *mean (μ) zero and standard deviation one (σ).*
- On the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.
  Formula: $\mathbf{X_{new} = (X_i - \mu) / \sigma}$

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

Ans:

VIF - the variance inflation factor -The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. (VIF) =1/(1-R_1^2). If there is perfect correlation, then VIF = infinity. Where R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables- If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it’s R-squared value will be equal to 1.So, VIF = 1/(1-1) which gives VIF = 1/0 which results in “infinity”.

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

Ans:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.It is used to compare the shapes of distributions. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

The q-q plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behavior?