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. There are following categorical variables in the dataset, 'season', 'yr', 'mnth', 'holiday', 'workingday', 'weekday', 'weathersit'. These variables were plotted against dependent variable (‘cnt’) using box plot and below inferences can be drawn –

1. Season - Season: 3: Fall has highest demand for rental bikes and Season 1: Spring has the least demand

2. **Year** - The demand for the bikes has increased in the year 2019

3. Month - Demand is continuously growing each month till June. September month has highest demand. After September, demand is decreasing

4. Holiday - The demand is less on holidays

5. Working day and Week day - Weekdays and Working days are not showing a distinct pattern for demand

6. Weather situation - Clear weather has shown higher demand for rental bikes, while there is no demand in extreme weather of heavy snow and heavy rain

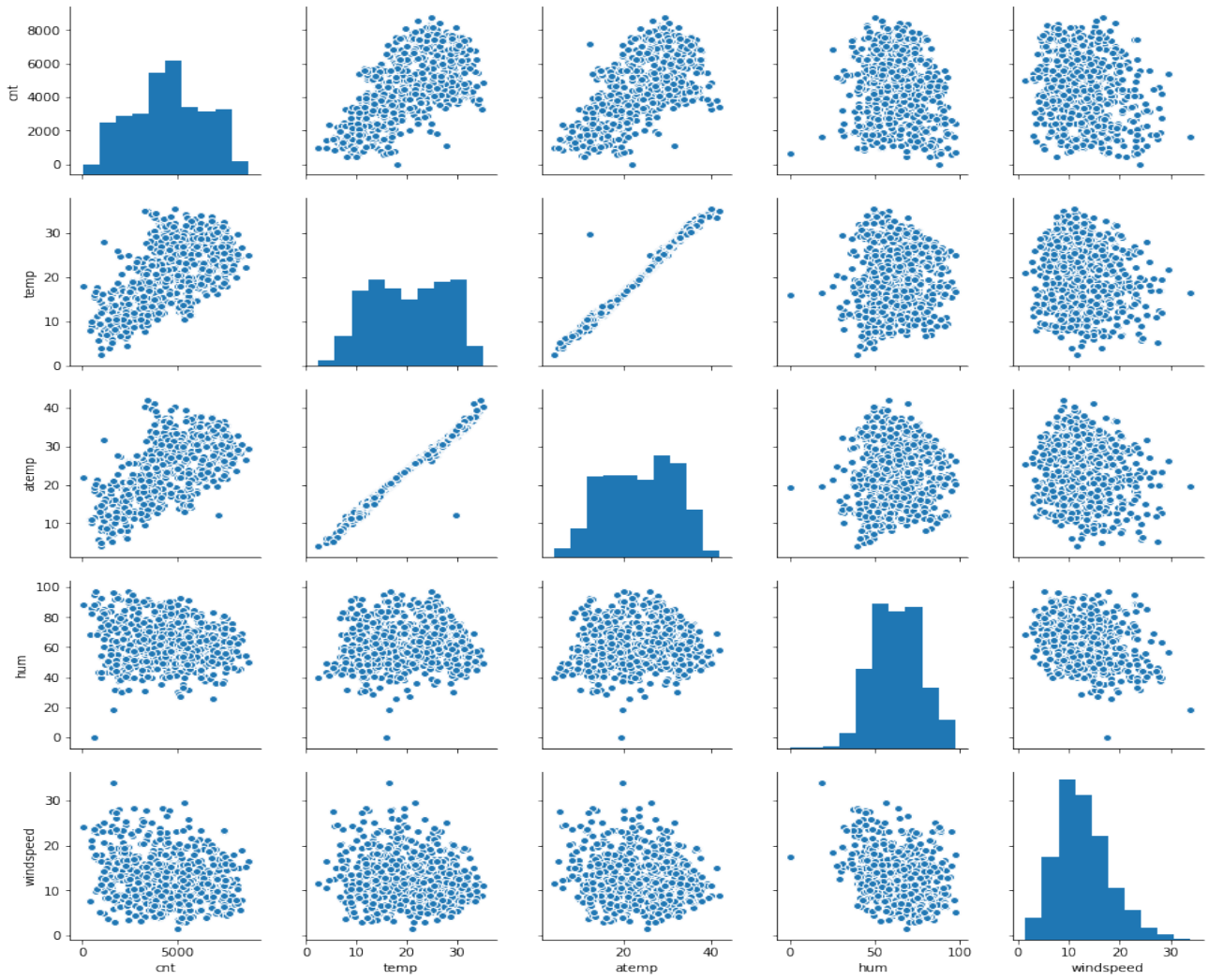
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Ans. It helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations (reduntant) created among dummy variables.

Another reason is, if we have all dummy variables it leads to Multicollinearity between the dummy variables, so drop_first = True is used

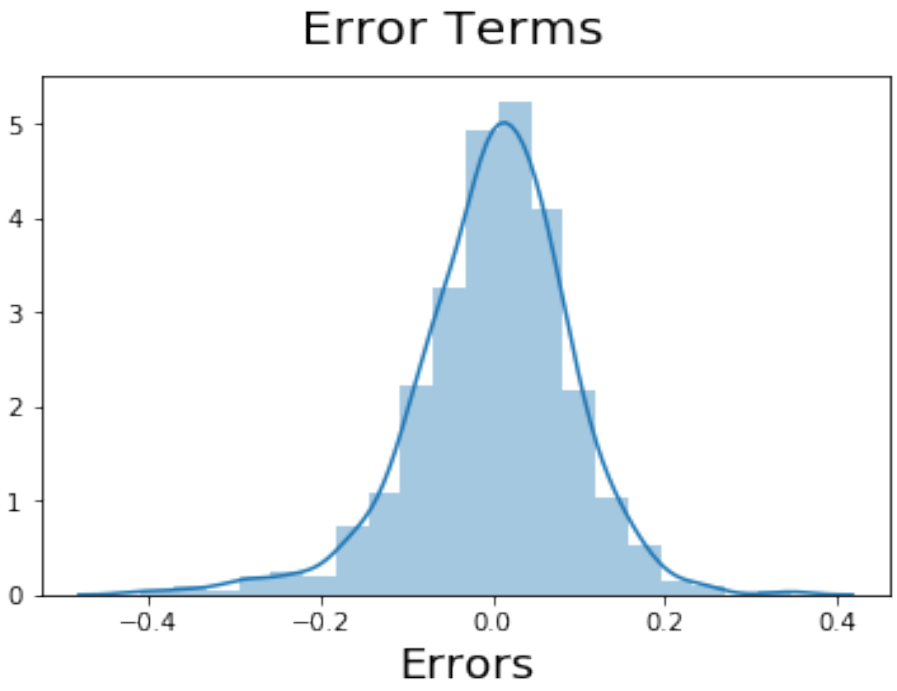
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans. When pair-plot was plotted among the numerical variables ‘atemp’ and ‘temp’ variables have the highest correlation with the target variable ‘cnt’.



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

Ans. Residuals distribution should follow normal distribution and centered around 0(mean = 0). We validated this assumption by plotting a distplot of residuals and verifying if residuals are following normal distribution or not. The below diagram shows that the residuals are distributed about mean = 0.



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

Ans. The top 3 features (predictor variables) that influences the bike booking significantly are:

1. Temperature (temp) - A coefficient value of '0.517673' indicated that a unit increase in temp variable increases the bike hire numbers by 0.517673 units.

2. Weather (weathersit_Light snow and rain) - A coefficient value of '-0.282580' indicated that, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.282580 units.

3. Year (yr) - A coefficient value of '0.233114' indicated that a unit increase in yr variable increases the bike hire numbers by 0.233114 units.

General Subjective Questions

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

Ans. Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values.

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

1. Simple Linear Regression: SLR is used when the dependent variable is predicted using only one independent variable.

2. Multiple Linear Regression: MLR is used when the dependent variable is predicted using multiple independent variables.

The equation for MLR will be:

$Y=\beta 0+\beta 1X1+\beta 2X2+...+\beta pXp+\epsilon$

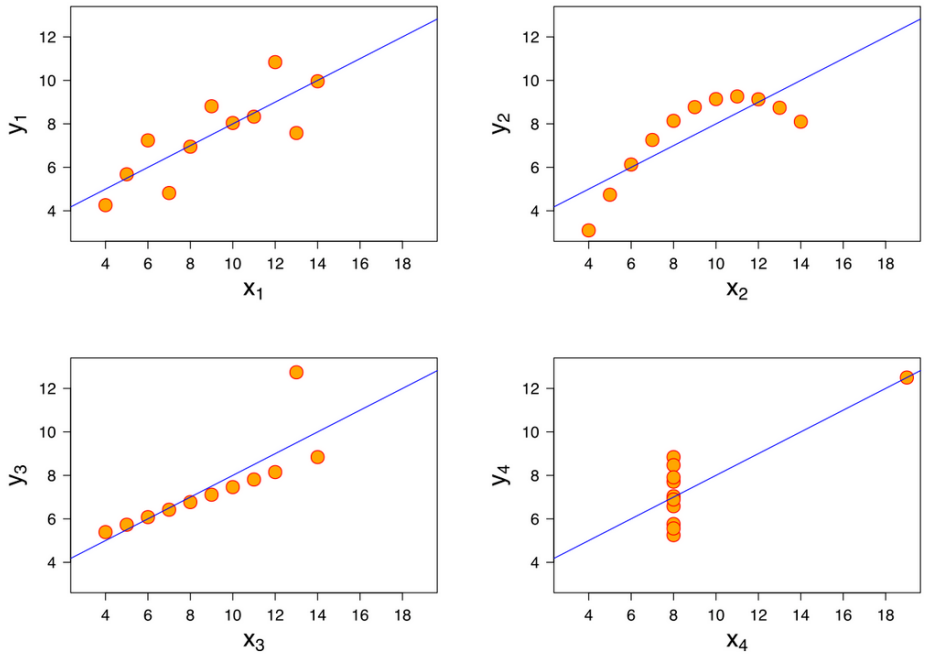
2. Explain the Anscombe's quartet in detail. (3 marks)

Ans. Anscombe's quartet was developed by statistician Francis Anscombe to demonstrate the importance of graphical analysis and the effect of outliers and influential observations when analyzing a dataset. It comprises of four datasets with same statistical features but when plotted on a graph, they have very different distributions and looked totally different.

Data – Statistical Properties for all four datasets

Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x* : s^2	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y* : s^2	4.125	±0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear Regression line	*y* = 3.00 + 0 .500*x*	to 2 and 3 decimal places, respectively
Coefficient of Determination of the linear regression	0.67	to 2 decimal places

Graphs –



Graphical Analysis –

1. The first scatter plot (top left) appears to be a simple linear relationship.
2. The second graph (top right) is not distributed normally; but there is a relation between them which is not linear.

3. In the third graph (bottom left), the distribution is linear, but should have a different regression line. There is an outlier present
4. The fourth graph (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. Correlation between sets of data is a measure of how well they are related. The most common measure of correlation in stats is the Pearson Correlation. The full name is the Pearson Product Moment Correlation (PPMC). It shows the linear relationship between two sets of data. Pearson's R measures the strength of the linear relationship. In simple terms, it answers the question, Can I draw a line graph to represent the data?

Pearson's r is always between -1 and +1

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 step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range.

Scaling of variables is an important step because if not done then algorithm weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values. Scaling just affects the coefficients of the model.

It also helps in speeding up the calculations in an algorithm.

The difference between Normalization and Standardization –

Normalization (Min Max Scaling) –

Normalization typically means rescales the values into a range of [0,1].

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

It is really affected by outliers.

Scikit-Learn provides a transformer called MinMaxScaler for Normalization.

It is useful when we don't know about the distribution

Standardization –

Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

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

It is much less affected by outliers.

Scikit-Learn provides a transformer called StandardScaler for standardization.

It is useful when the feature distribution is Normal or Gaussian.

It is a often called as Z-Score Normalization.

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^2).

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables 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. Q Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile.

It is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

In linear regression its importance lies to check when we have training and test data set received separately then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behavior