Submitted by:
Kiran Kumar Valluru (APFE21709855)

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 categorical variable in the dataset are 'season', 'weathersit', 'holiday', 'mnth', 'yr' and 'weekday'. These have been visualized and inferred as follows:

- The demand for bike is less in "Spring" season when compared with other seasons and "Fall" season has high demand.
- The demand for bike increased in the year 2019 when compared with year 2018.
- Month Jun to Sep is the period when bike demand is high with September being highest. The Month Jan is the lowest demand month.
- Bike demand is less in holidays in comparison to not being holiday.
- The demand for bike is almost similar throughout the weekdays.
- There is no significant change in bike demand with working day and non-working day.
- The bike demand is high when weather is "Clear" and less in case of "Light Snow and Light Rain". We do not have any data for "Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog" , so we cannot derive any conclusion. It could be that the company is not operating on those days or there is no demand of bike.

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

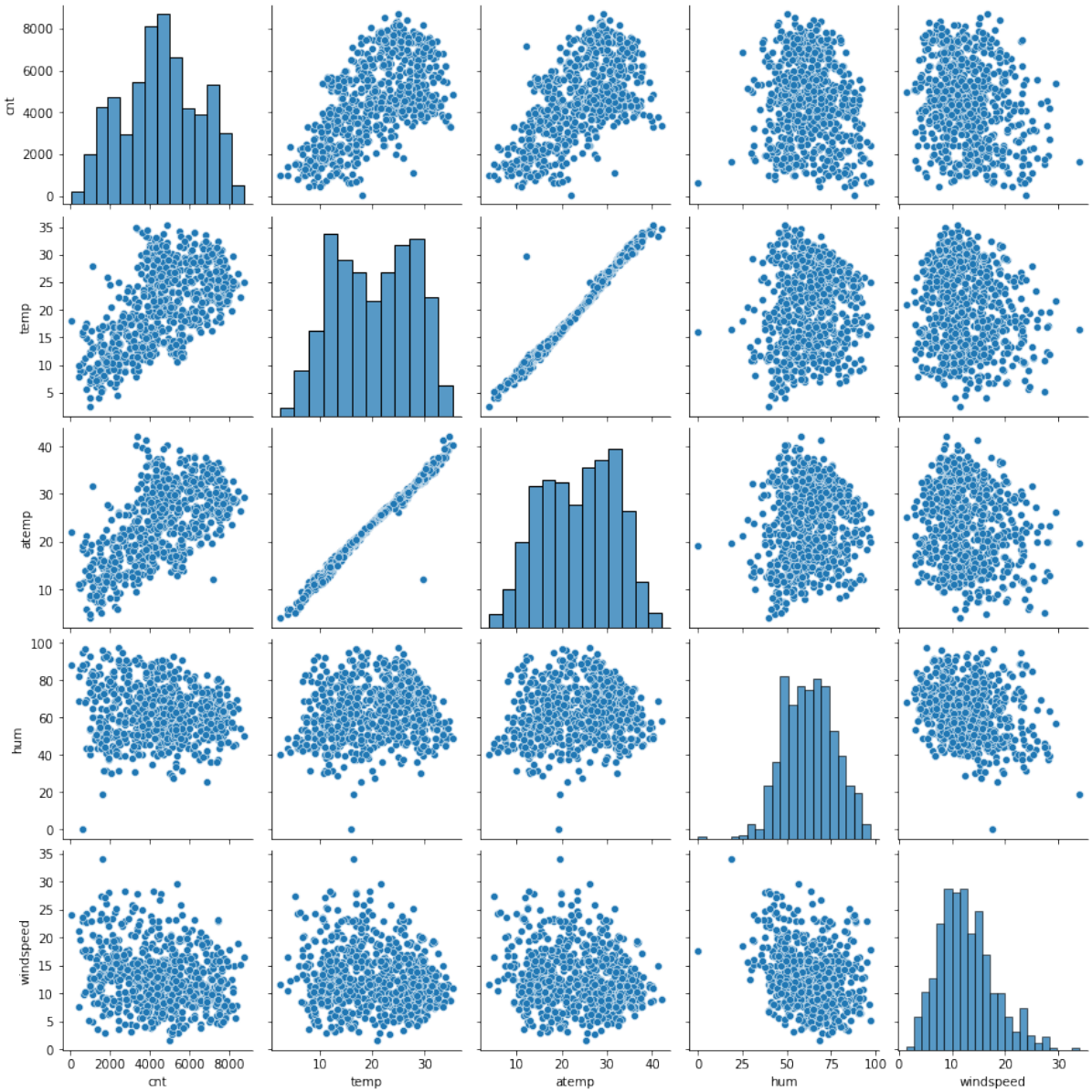
Answer:
'drop_first=True' is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables (Dummy variable trap).

Let's say we have 3 types of values in Categorical column (A, B & C) and we want to create dummy variable for that column. If one variable is not A and B, then It is obvious C. So, we do not need 3rd variable to identify the C.

If we don't drop the first column then our 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's may be distorted.

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

Answer:

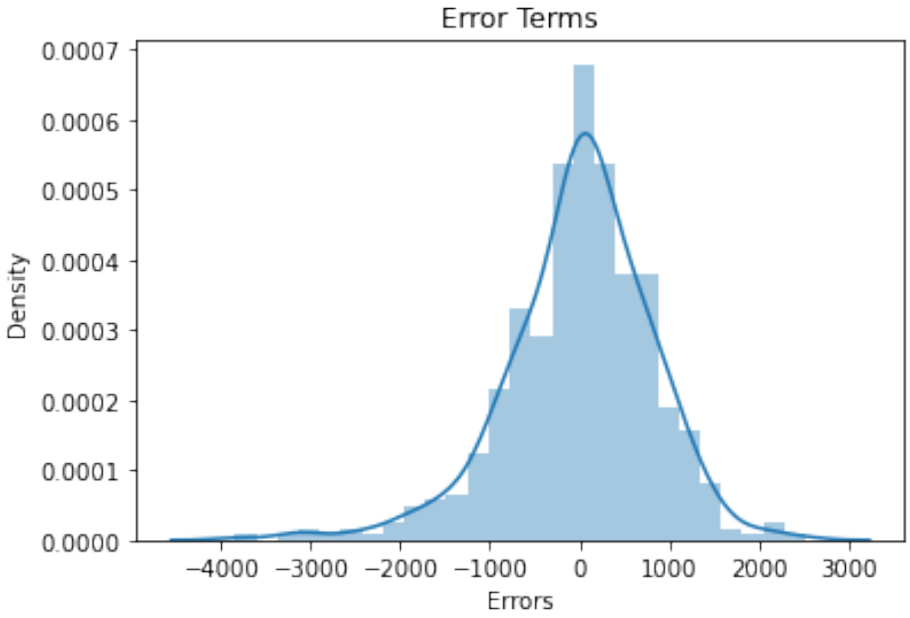


From the above pair plot, ‘temp’ and ‘atemp’ are the two numerical variables which are highly correlated with the target variable 'cnt'

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

Answer:
In order to validate assumptions of the model, and hence the reliability for inference, we go with the following procedures:

- **Residual Analysis:**
  We need to check if the error terms are normally distributed. Residuals distribution should follow normal distribution and centred around 0 (mean = 0). I have validated this assumption about residuals by plotting a histogram of residuals and see if residuals are following normal distribution or not. The below plot shows that the residuals are distributed about mean = 0.



- **There is a linear relationship between X and Y**
  Using the pair plot, we could see there is a linear relation between 'temp' and 'atemp' variable with the target 'cnt'.

- **There is No Multicollinearity between the predictor variable**
  From the VIF calculation we could find that there is no multicollinearity existing between the predictor variables, as all the values are within permissible range of below 5

- **No Heteroskedasticity**
  There's no funnel like pattern in the Residual vs Fitted values plot, so no heteroskedasticity present.

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

Answer:
The top 3 features significantly contributing towards demand of shared bikes are:

1. Temperature (temp) -> Positive
2. Weather Situation (weathersit_Light Snow & Rain) -> Negative
3. Year (yr) -> Positive

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:
Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

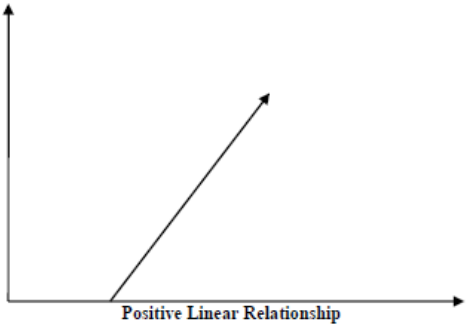
Mathematically the relationship can be represented with the help of following equation:

$$Y = mX + b$$

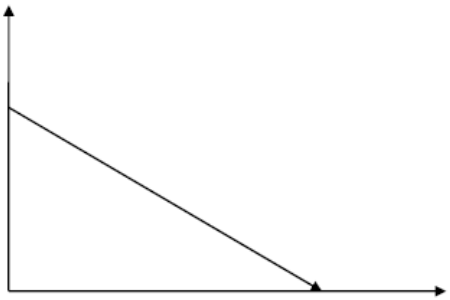
Here,
Y is the dependent variable we are trying to predict.
X is the independent variable we are using to make predictions.
m is the slop of the regression line which represents the effect X has on Y
b is a constant, known as the Y-intercept. If X = 0, Y would be equal to b.

Furthermore, the linear relationship can be positive or negative in nature as explained below:

Positive Linear Relationship
A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph:



Negative Linear relationship
A linear relationship will be called negative if independent increases and dependent variable decreases. It can be understood with the help of following graph:



Negative Linear Relationship

Linear regression is of the following two types:

- Simple Linear Regression
- Multiple Linear Regression

Assumptions
The following are some assumptions about dataset that is made by Linear Regression model:

Multi-collinearity − Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.

Auto-correlation − Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

Relationship between variables − Linear regression model assumes that the relationship between response and feature variables must be linear.

Normality of error terms - Error terms should be normally distributed.

Homoscedasticity - There should be no visible pattern in residual values.

2. Explain the Anscombe's quartet in detail

Answer:
Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change **completely**

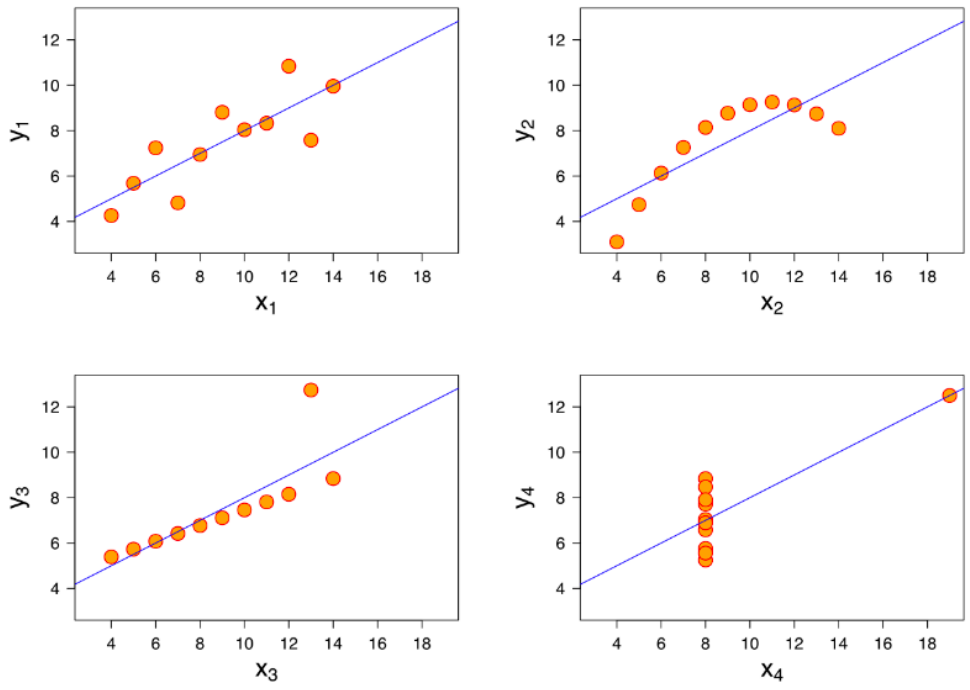
when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

	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

The summary statistics show that the means and the variances were identical for x and y across the groups:

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset.
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset.

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines aswell, but each dataset is telling a different story:



- Dataset 1 appears to have clean and well-fitting linear models.

- Dataset 2 is not distributed normally.
- In Dataset 3 the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset 4 shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

3.What is Pearson's R?

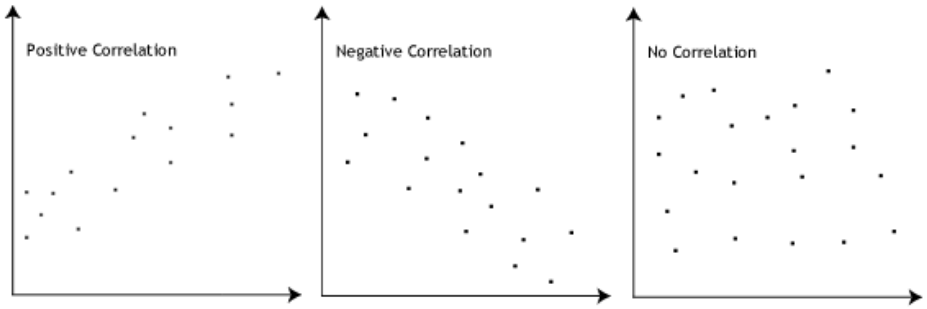
Answer:

The Pearson's correlation, also known as Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association

This is represented in the below figure:



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer:

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying

magnitudes or values or units. 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, regardless of the unit of the values.

Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and, in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

Normalization or Min-Max Scaling is used to transform features to be on a similar scale. This scales the range to [0, 1] or sometimes [-1, 1].

Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.

Difference between Normalization and Standardization:

S. No	Normalization	Standardization
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is often called as Scaling Normalization	It is often called as Z-Score Normalization.

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

Answer:

VIF - the variance inflation factor - The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity.

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) => infinity. To solve this, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Answer:

The Quantile-Quantile (Q-Q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Usage:

- It can be used with sample sizes aswell
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

The Quantile-Quantile plot is used for the following purpose:

- Determine whether two samples are from the same population.
- Whether two samples have the same tail
- Whether two samples have the same distribution shape.
- Whether two samples have common location behaviour.

Importance of Q-Q plot:

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.