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:

Insights obtained from plotting the categorical columns against the target are as follows.

From this plot we can see a few trends:

- Season 1 (Spring Season) has the least count compared to the others ((1: spring, 2: summer, 3: fall, 4: winter). This could affect our predictions
- we can see a trend with the months as well, this may be since seasons and weather are linked to the month. This could affect our predictions
- clearly the count is much lesser in Snow/Rain environment conditions. This could affect our predictions
- Our data set says that the rental is more in 2019 (This may be due to the popularity it gained over the previous year) , which could also help with the predictions

There are certain patterns that exist between the categorical columns and the target that can help predict the target.

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

Answer:

Drop_first=true is a parameter that we use in pd.get_dummies.

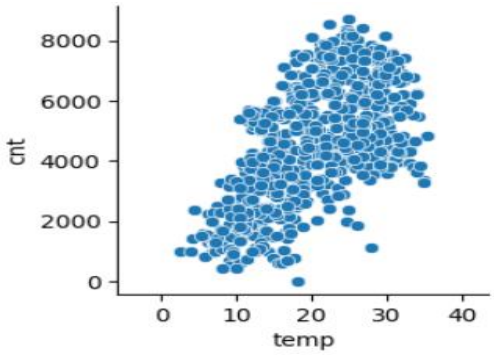
We do this to avoid a phenomenon called dummy variable trap. This trap happens when all dummy variables are correlated and when we are using a linear model.

It will violate the assumption that multicollinearity does not exist, thus giving co-efficient that are not possible to interpret and become unreliable.

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

Answer:

Looking at the pair plot, we can see that temp and atemp have a linear relation with the target variable. However, using temp and atemp in our model together may not yield the best results as they contain values that are similar in nature hence leading to a multicollinearity effect

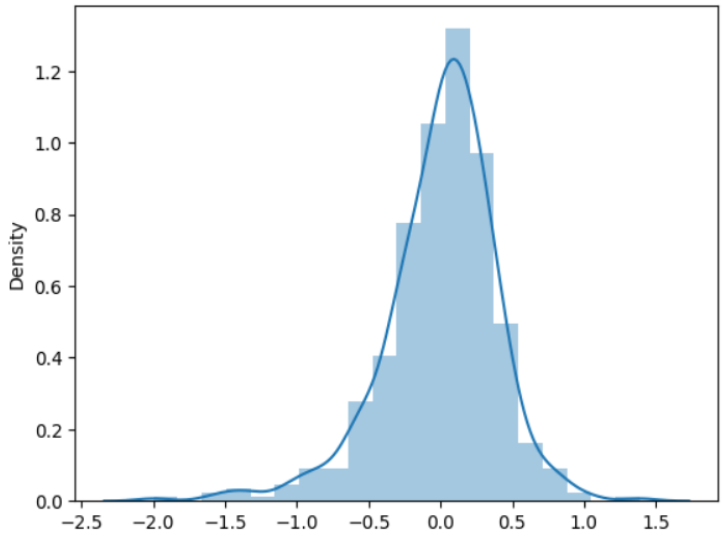


4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

Answer:

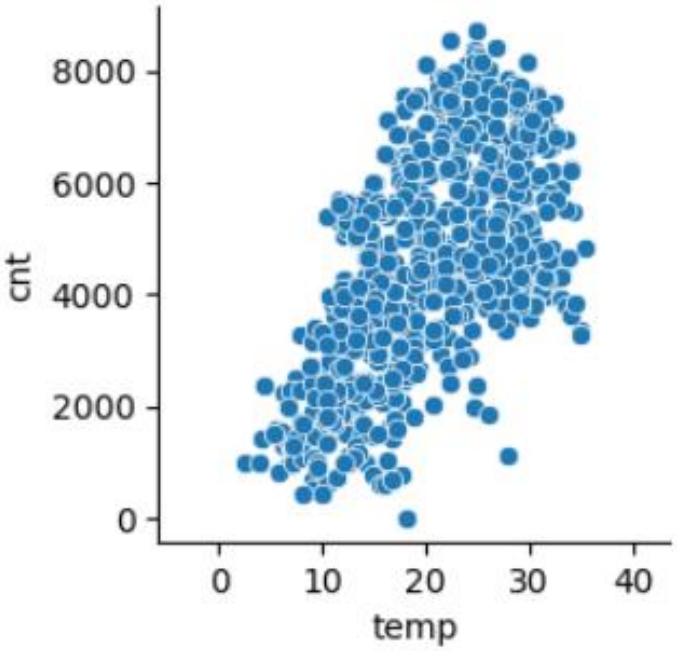
- **Ensuring the residuals are normally distributed**:

We have plotted the distplot of the residuals and we can see that it has taken a bell shape normal distribution curve.



- **Linear relationship between X and y**:

On making a pair plot during Bi-Variate analysis we can see that there is a linear relationship between Target (**cnt**) and variables **temp** and **atemp**.



- **There is no multicollinearity between the predictors:**

Once the model was built with all the significant predictors (p-value < 0.05) , we run a VIF function. We noticed that the VIF value for all predictor value are <=5 , 5 being the cut off value.

	Features	VIF
0	const	14.77
5	season_spring	4.14
2	temp	3.51
6	season_winter	3.47
3	hum	1.99
9	mnth_4	1.68
1	workingday	1.65
15	weekday_6	1.64
17	weathersit_Misty	1.60
13	mnth_10	1.54
10	mnth_5	1.40
11	mnth_6	1.27
16	weathersit_Light Snow/Rain	1.25
4	windspeed	1.22
12	mnth_9	1.19
8	mnth_3	1.16
14	weekday_1	1.05
7	yr_2019	1.04

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:

1) Yr – if the yr value is 2019 it contributes to 1.0239 increase in the count.
2) weathersit_Light Snow/Rain – if the weather is of this type then there is a decrease in demand by -1.2188
3) temp – every unit increase in temp raises the count by 0.4649.

General Subjective Questions

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

Answer:

Linear regression is a statistical technique for modelling the relationship between a dependent variable and one or more independent variables. Simple linear regression is the most basic method, which uses one independent variable to predict a dependent variable by fitting a linear equation to observed data.

The equation is in the form:

y = B0 + B1(X1) + E

Where:

y is the dependent variable.

X1 is the independent variable.

B0 is the intercept.

B1 is the slope of the line.

E is the error term.

The goal is to find the best-fitting line, which minimizes the sum of squared residuals (the differences between observed and predicted values). This is typically done using the least squares method. Linear regression assumes a linear relationship, homoscedasticity (constant variance of errors), independence of errors, and normally distributed errors.

For multiple linear regression, the concept extends to multiple independent variables:

y = B0 + B1x_1 + B2x_2 + ...+ \B(n)x_n + E

Here, the algorithm estimates the coefficients B1, B2,...B(n) that best predict y based on the input features.

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

Answer:

Anscombe's quartet is a collection of four datasets with practically similar simple descriptive statistics (mean, variance, correlation, and linear regression line), yet when plotted, they look drastically different. This was established by Francis Anscombe in 1973 to emphasize the significance of visualizing data before analysing it. The four datasets show that depending exclusively on statistical summaries can be deceiving; visual study of data can reveal patterns, trends, and anomalies that statistics alone cannot.

Each dataset in the quartet contains:

- The same mean for both x and y variables.

- The same variance.

- The same correlation coefficient (Pearson's r).

- The same regression line.

Yet, the scatterplots show different relationships between the variables (linear, non-linear, outliers, etc.), underscoring the value of data visualization in statistical analysis.

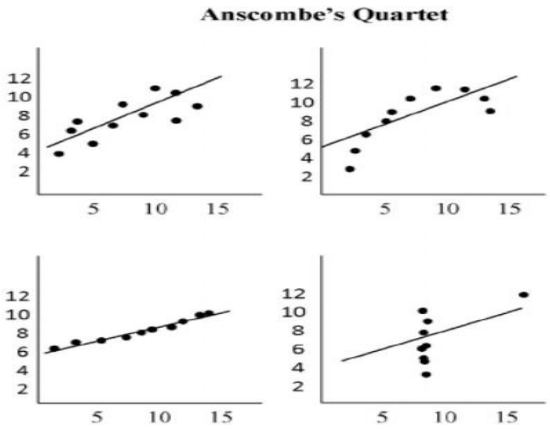
Example: Based on the Data points, numerically they look the same on paper.

Data sets for the 4 XY plots

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	5.76
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	8.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	7.26	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Property	Value
Mean of X (average)	9 in all 4 XY plots
Sample variance of X	11 in all four XY plots
Mean of Y	7.50 in all 4 XY plots
Sample variance of Y	4.122 or 4.127 in all 4 XY plots
Correlation (r)	0.816 in all 4 XY plots
Linear regression	y = 3.00 + (0.500 x) in all 4 XY plots

But when we plot the values, we see the following:



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

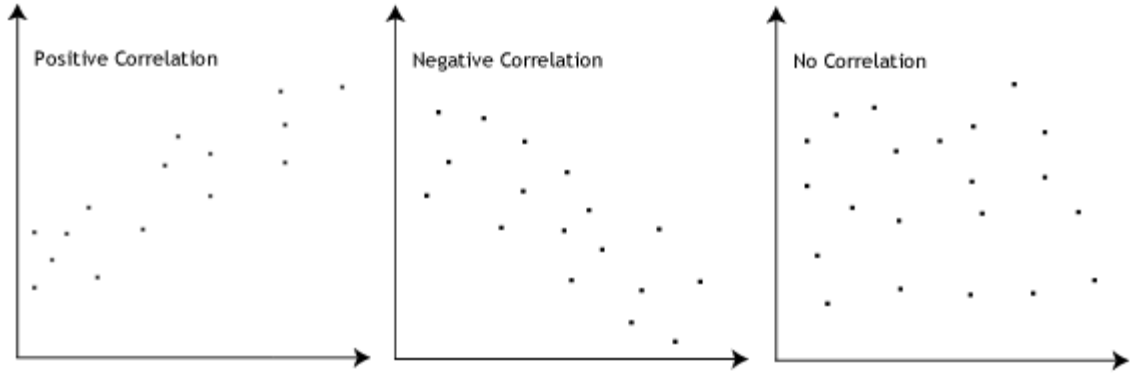
Answer:

Pearson's R, or Pearson's correlation coefficient, measures the strength and direction of the linear relationship between two variables. It is a value between -1 and 1, where:

+1 indicates a perfect positive linear relationship.

-1 indicates a perfect negative linear relationship.

0 indicates no linear relationship.



Pearson's R is computed by dividing the covariance of the variables by the product of their standard deviations. It is a common statistical metric that is particularly effective when determining the linear relationship between two continuous variables. However, it can only detect linear correlations and is sensitive to outliers.

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 of adjusting the range of feature values to a standard range, typically between 0 and 1 or according to a standard normal distribution.

It is performed to ensure the following:

- To ensure that features contribute equally to the model.
- To speed up convergence in optimization algorithms like gradient descent.
- To prevent features with larger scales from dominating others.

Difference Between Normalized Scaling and Standardized Scaling:

Normalized Scaling (Min-Max Scaling): Transforms features to a fixed range, typically [0, 1]. It is sensitive to outliers and is used when you want to preserve the relationships between features while constraining their range.

X (normalised value) = [X – X(min)]/[X(max)-X(min)]

Standardized Scaling (Z-Score Normalization): Centres the data around the mean with a unit standard deviation. It is useful when features have different units or ranges, as it transforms them to a standard normal distribution (mean = 0, variance = 1).

X (standarised value) = **[**X – mean ((values of the column))**]** / Std Deviation (values of the column)

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

Answer:

The Variance Inflation Factor (VIF) estimates how much a regression coefficient's variance is inflated because of model multicollinearity. A VIF value of infinity is often obtained when there is perfect multicollinearity, which means that one independent variable is a perfect linear combination of the other variables in the model. This results in a division by zero in the VIF calculation, as the determinant of the correlation matrix (which is part of the VIF formula) is zero.
An infinite VIF indicates that the regression model is unable to distinguish between fully correlated predictors, making the estimations untrustworthy. It emphasizes the need to eliminate or combine strongly collinear variables.

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

Answer:

A Q-Q plot (Quantile-Quantile plot) is a graphical tool for comparing the distribution of a dataset to a theoretical distribution, typically the normal distribution. It compares the quantiles of the sample data with the quantiles of the theoretical distribution. If the data is regularly distributed, the points on the Q-Q plot will roughly follow a straight line.

Use and Importance of Linear Regression:

Assessing the Normality of Residuals:

Linear regression presupposes that the residuals (errors) follow a normal distribution. A Q-Q graphic can assist verify this assumption. Deviations from the straight line show non-normality, which could point to model issues, such as the necessity for a dependent variable transformation.

Identifying Outliers:

Outliers appear as points that deviate greatly from the line and may require further research. Q-Q plots are essential for evaluating the assumptions behind linear regression and ensuring the model's inferences and predictions are correct.