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 1:

Categorical variables are observed to have an impact on dependent variable 'Count' of shared bikes opted in a day and hence we can say that demand is impacted by categorical variables.

Some categorical variables have positive correlation; some have negative correlation.

Some categorical variables like workingday_1, weekday_1-weekday_5 and holiday_1 were found to have infinite VIF, i.e. all the variance in these variables is already explained by other independent variables and hence these features were excluded from the model.

In the final model season_2, yr_1, mnth_9, mnth_10, mnth_11, weekday_6, weathersit_3 features were included in the final model which predicts the value of dependent variable.

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

Answer 2:

A categorical variable with n unique values can be explained by n-1 dummy variables which are one hot encoded. Since one value out of n distinct values can be identified if rest of n-1 values are 0.

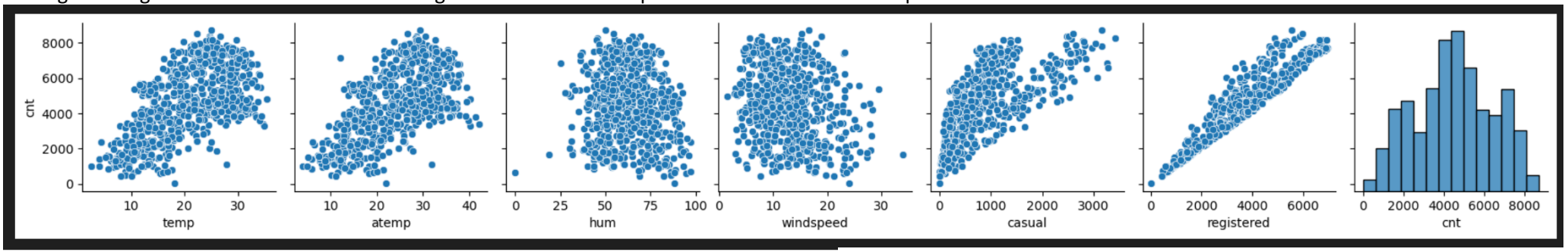
For example, If we have three values: Ture, False, Maybe for a category 'test', and we encode these values by creating three dummy columns test_True, test_False and test_Maybe. If for a row test_True and test_False are 0 then we already know that test has value Maybe for that row and hence we do not require the third dummy column.

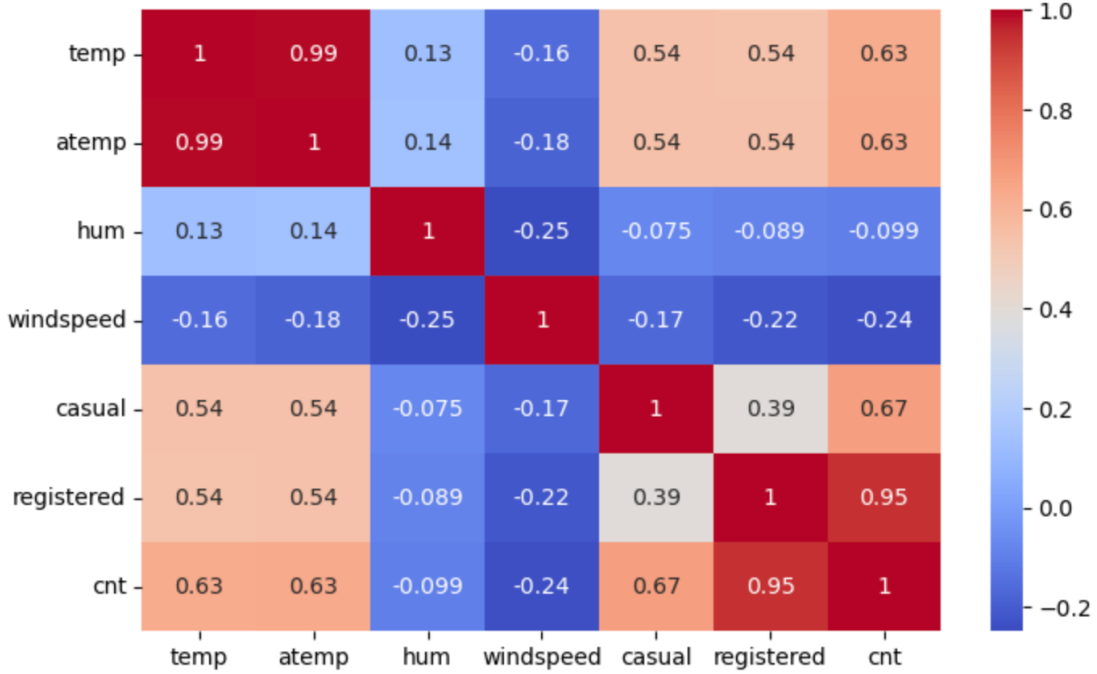
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable

Answer 3:

Looking at the pair plot of numerical variables ‘registered’ seemed to have highest correlation with count, but since registered is kind of what dependent variable cnt is made up and semantically Is almost the demand itself.

If we ignore ‘registered’ and ‘casual’ then the highest correlation of dependent variable is with ‘temp’

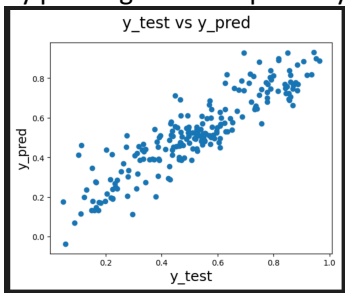




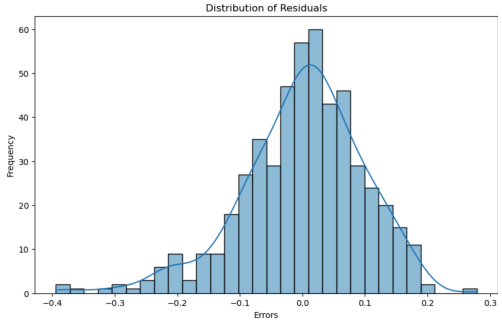
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer 4:

Assumptions of linear regression are:

- Linear relation between independent and dependent variable:
  - Before building the model checked the pairplot (numeric feature vs dependent variable), correlation matrix (heatmap)
  - After building the model R2 score paired with Residuals (Ytrain-Ypred) distribution
- Homoscedasticity (Constant variance in error)
  - By plotting a scatter plot of ytest vs ypred, which shows even spread of point around line passing through origin
  - 

- Normal distribution of residuals with mean error(residual) at 0



- No multicollinearity (Association of predictor variables): Checked by using VIF (Variance Inflation factor)

	Feature	VIF
1	casual	5.28
0	atemp	4.73
3	yr_1	2.02
2	season_2	1.54
7	weekday_6	1.46
4	mnth_9	1.24
5	mnth_10	1.21
8	weathersit_3	1.12
6	mnth_11	1.09

- Independence of error terms

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

Answer 5:

Top three features significantly contributing twowards the demand are ‘atemp’, ‘casual’ and ‘season2’ (followed by ‘yr_1’, ‘waethersit3’)

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0882	0.013	6.569	0.000	0.062	0.115
atemp	0.4260	0.027	15.839	0.000	0.373	0.479
casual	0.3440	0.032	10.827	0.000	0.282	0.406
season_2	0.0305	0.011	2.722	0.007	0.008	0.052
yr_1	0.2059	0.009	22.317	0.000	0.188	0.224
mnth_9	0.0884	0.017	5.145	0.000	0.055	0.122
mnth_10	0.0859	0.017	5.158	0.000	0.053	0.119
mnth_11	0.0830	0.016	5.084	0.000	0.051	0.115
weekday_6	-0.0559	0.014	-3.977	0.000	-0.083	-0.028
weathersit_3	-0.1983	0.027	-7.361	0.000	-0.251	-0.145

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

Linear regression algorithm helps to predict the values of a continuous target variable based on one or more input features.
It works on the following assumptions:

- **Linearity**: There should be linear relationship between dependent variable and the feature variables (independent variables)
- **Independence of Errors**: The error terms found in prediction should be independent and should not have any relation or pattern
- **Normal distribution of Error:** Errors should be normally (naturally) distributed with mean at 0
- **Homoscedasticity:** Error terms have constant variance

Algorithm works by assuming that the value of target variable (dependent variable) 'y' can be determined using equation of straight line,
i.e. "y = mX + c".

Then the algorithm with the help of training data set tries to predict the value of target variable 'ypred' such that the error "y-ypred" can be minimized.
A cost function is based on this fact and then algorithm tries to minimise this cost function.

Minimization of cost function is done either using 'Ordinary least sum of squares' or 'Gradient Descent'.
The coefficients in equation of line (m, in y=mX+c) calculated such that the cost function is minimum.

2. Explain the Anscombe's quartet in detail.

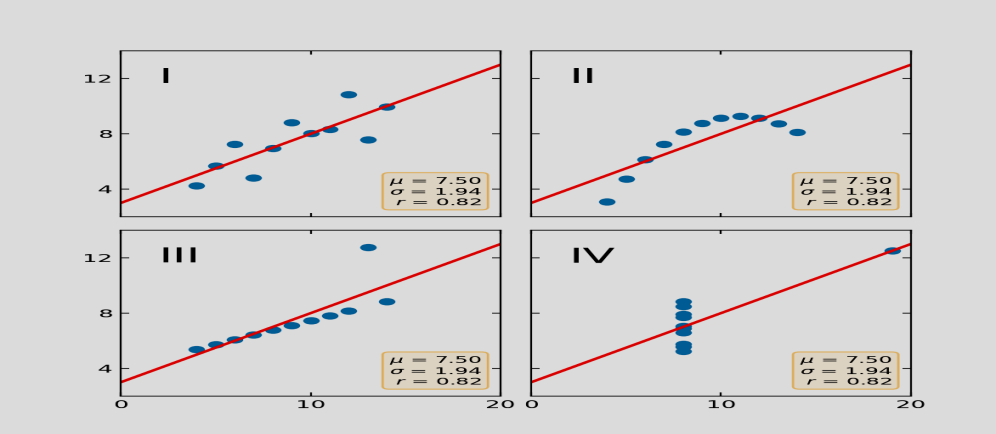
Answer:

Anscombe's quartet is special example of four data set which when visualised using plots like scatter plots appears to be very different but statistical measures like Mean, Standard deviation etc are identical.

Even the linear regression lines of the plot are identical.

It is often used to emphasize on the importance of visualization rather than just looking at the data statistically.

Although statistically similar the data is qualitatively different.



3. What is Pearson's R?

Answer:

Pearson's R is measure of linear correlation between to attributes of data.

The value ranges between -1 and +1, values near +1 indicates strong positive correlation, values near -1 indicates strong negative correlation and values near 0 indicates no linear relation

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

Answer:

Sometimes it may happen that different variables in the data set have different range of values, for example in dataset of house rooms can be in range of 1-10 whereas the variable price could range from 5000000 to 100000000 due to which the visualization or statistical comparison of variables depicting their relationship might not be very precise and may miss out some patterns.

This could be resolved by scaling whereby the means of Standardisation or Normalisation the values of these different variables can be brought to a comparable range such that the relationship between them can be analysed more clearly.

Parameter	Normalized Scaling (Min-Max Scaling)	Standardized Scaling (Z-score Scaling)
Definition	Scales data to a fixed range, typically **0 to 1**.	Transforms data to have a **mean of 0** and a **standard deviation of 1**.
Formula	$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$	$X_{\text{std}} = \frac{X - \mu}{\sigma}$ (μ=mean, σ=std dev)
Outlier Sensitivity	**Highly sensitive** to outliers.	**Less sensitive** to outliers.
Distribution	Makes no assumption about the data's distribution.	Assumes a Gaussian (normal) distribution.

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

Answer:

Infinite value of VIF suggest perfect multi-collinearity among independent variables, which in simple term means that one independent variable can be predicted by other independent variables.

For example, there are two features age in months and age in years, which effectively means age and hence are redundant. If both features are kept in the model it could lead to inflated error or reduced model efficiency.

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

Answer:

A Q-Q plot is a graphical method to compare two probability distributions by plotting their quantiles against each other to see if they come from the same distribution.
A Q-Q plot (Quantile-Quantile plot) is a scatter plot that compares the quantiles of a sample data distribution against the quantiles of a theoretical distribution.
The quantiles of your data are plotted on one axis, and the corresponding quantiles of a theoretical distribution (like the normal distribution) are plotted on the other.

The primary use is to check if the **residuals** of a linear regression model are normally distributed. This is a fundamental assumption for many linear regression tests to be valid.

The plot can reveal outliers, as points far from the line may represent unusual errors.\

A Q-Q plot provides a quick visual check that is often more conclusive than a histogram for assessing normality. It helps a modeler decide if the linear regression model is appropriate and whether the model's results are reliable, guiding further analysis or model adjustments.