Assignment-based Subjective Questions

Q1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

The effect of categorical variables on the dependent variable can often be assessed through statistical tests , visualisations, and regression analysis. These methods help determine whether there are significant differences in the dependent variable across different categories of the categorical variable.

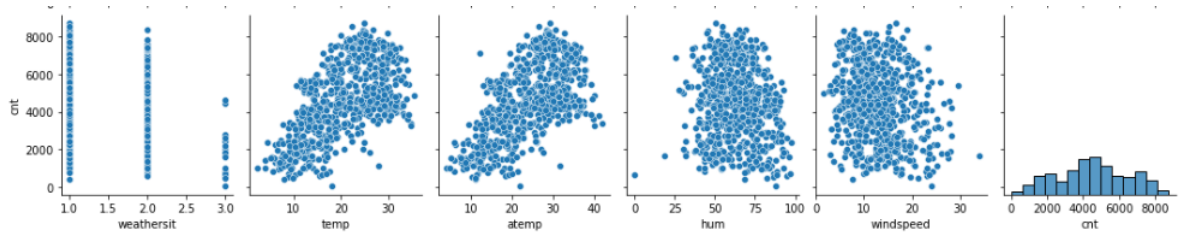
- May to October has higher average Demand for cycles
- On an average more usage was there on working days compared to holidays(considering 1 is a holiday in that column) , summer was an exception.
- Average demand for cycles remained higher demand for cycles in Summer and Fall
- Clear Skies showed more demand compared to rainy Days
- Irrespective of season , rainy days saw less demand compared to clear/cloudy skies
- Weekdays 4& 5 Saw more demand for cycles consistently during Spring
- During Summer the average demand for cycles is almost similar on most of the weekdays
- During Fall WeekDay2 saw more Average Demand for cycles compared to any other day

Q2: Why is it important to use drop_first=True during dummy variable creation?

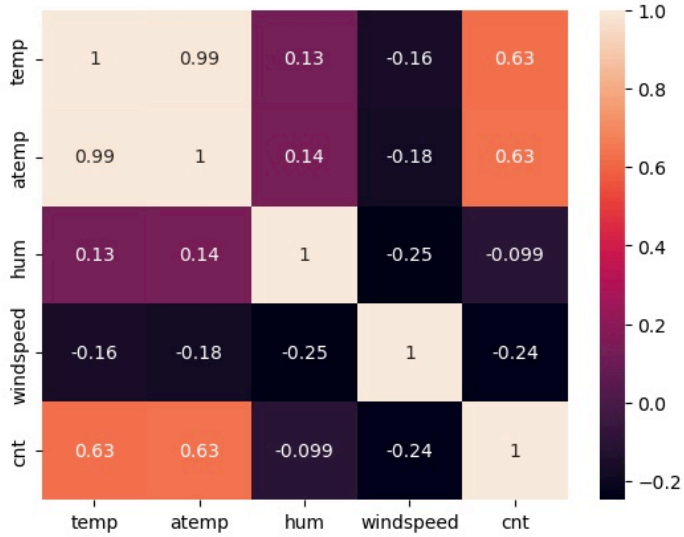
When creating dummy variables from categorical variables, it's important to use the drop_first=True parameter to avoid multicollinearity issues in regression analysis, specifically in situations where the categorical variable has more than two levels. For example, if you have dummy variables for "Male", "Female", and "Other", knowing the values of "Male" and "Other" would automatically determine the value of "Female".ncluding all dummy variables without dropping one can lead to redundancy and unnecessary complexity in the model. Dropping the one level helps to simplify the model while preserving the necessary information about the categorical variable.

Q3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

This pair-plot helps visualize potential relationships and patterns in the data. From below screenshot that shows a relationship between some of the variables we can observe clear concentration of datapoints of cat vs temp & atemp indicating positive correlation while with the other numerical attributes like humidity and windspeed not that much of a positive correlation as the scatter plot seems to have data concentration around a certain range. We can clearly observe from the next heat map the clear values of co-relation .




```
08]: sns.heatmap(df[num_cols].corr(),annot=True)
     plt.show()
```




```
[ ]: # From Above we can understand that temp and atemp are corelated in the same way with cnt
     # so we can consider this as one of the items for multicolinerity
```

Data Preparation for Linear Regression using RFE

As per above correlation matrix, target/dependent variable ‘cnt’ is positively correlated with ‘temp’, ‘atemp’ and negatively correlated with ‘humidity’, ‘wind speed’ numerical independent variables.

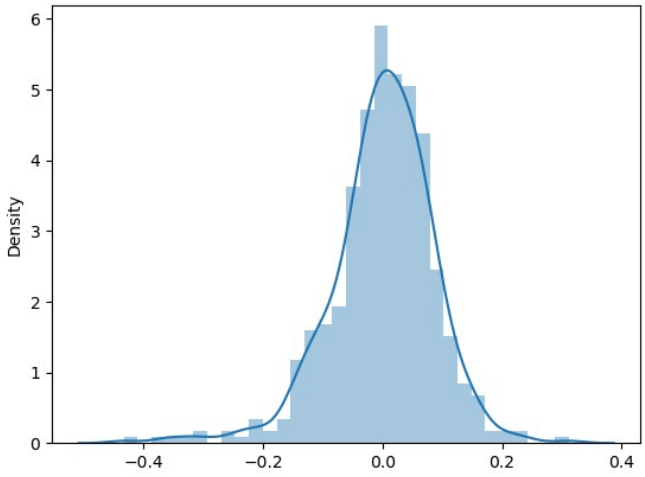
Q4: How did you validate the assumptions of Linear Regression after building the model on the training set?

1. **Residual Analysis**: error terms are normally distributed from the residual analysis.

   ○

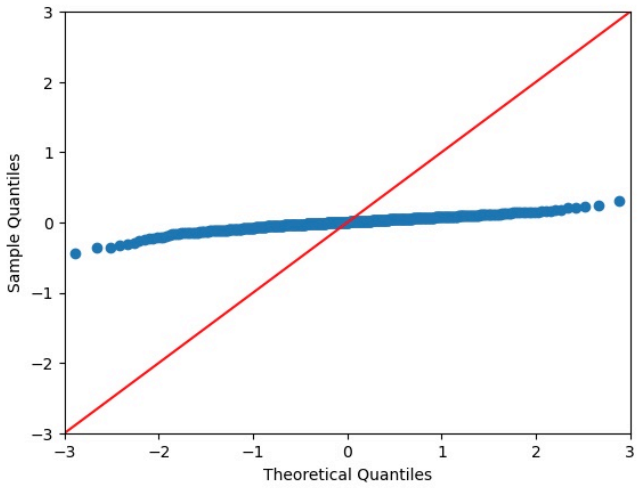
Residual Analysis

```
73]: y_train_pred = lm.predict(x_tran_new)

74]: res = y_train - y_train_pred
     sns.distplot(res)

74]: <Axes: ylabel='Density'>
```



2. **Error terms are centred around 0** :

```
8]: sm.qqplot(res,line='45')

8]:
```



3. **Multicollinearity**:

   - Variance inflation factor (VIF) for each all variables in the model is less than 5 and p values are less than 0.05

```
vif = pd.DataFrame()
X = x_tran_new
vif['Features'] = X.columns
vif['VIF'] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
vif['VIF'] = round(vif['VIF'], 2)
vif = vif.sort_values(by = "VIF", ascending = False)
print(vif)
```

	Features	VIF
0	const	68.81
3	temp	2.96
7	winter	2.38
11	jan	2.29
4	hum	1.91
6	summer	1.84
10	feb	1.75
12	nov	1.74
2	workingday	1.66
14	sat	1.66
9	dec	1.65
15	cloudy	1.58
8	aug	1.47
13	sep	1.27
16	rainy	1.26
5	windspeed	1.20
1	yr	1.03

Q5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

```
print(lm.summary())
```

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.850
Model:	OLS	Adj. R-squared:	0.845
Method:	Least Squares	F-statistic:	174.1
Date:	Wed, 29 May 2024	Prob (F-statistic):	5.58e-191
Time:	10:23:15	Log-Likelihood:	521.70
No. Observations:	510	AIC:	-1009.
Df Residuals:	493	BIC:	-937.4
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2355	0.033	7.247	0.000	0.172	0.299
yr	0.2307	0.008	28.956	0.000	0.215	0.246
workingday	0.0516	0.011	4.788	0.000	0.030	0.073
temp	0.4665	0.030	15.625	0.000	0.408	0.525
hum	-0.1484	0.037	-3.994	0.000	-0.221	-0.075
windspeed	-0.1912	0.025	-7.551	0.000	-0.241	-0.141
summer	0.0819	0.012	6.630	0.000	0.058	0.106
winter	0.1445	0.014	10.332	0.000	0.117	0.172
aug	0.0467	0.016	2.897	0.004	0.015	0.078
dec	-0.0482	0.018	-2.665	0.008	-0.084	-0.013
feb	-0.0416	0.021	-2.004	0.046	-0.082	-0.001
jan	-0.0708	0.021	-3.388	0.001	-0.112	-0.030
nov	-0.0456	0.018	-2.479	0.014	-0.082	-0.009
sep	0.1040	0.016	6.403	0.000	0.072	0.136
sat	0.0602	0.014	4.321	0.000	0.033	0.088
cloudy	-0.0598	0.010	-5.772	0.000	-0.080	-0.039
rainy	-0.2549	0.026	-9.781	0.000	-0.306	-0.204

Omnibus:	84.109	Durbin-Watson:	2.028
Prob(Omnibus):	0.000	Jarque-Bera (JB):	225.847
Skew:	-0.813	Prob(JB):	9.08e-50
Kurtosis:	5.826	Cond. No.	21.0

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Based on the final model, features such as temperature, humidity, windspeed, season(winter) are highly significant and contributing most significantly in explaining the demand of shared bikes.Also clearly from the correlation table and graphs above this helps explain. Also we see that there are couple of other categorical variables which got included as part of the model which are negatively correlated such as ‘Rainy’ season which are coming out to be significant.

General Subjective Questions

Q1: Explain the linear regression algorithm in detail.

Linear regression is a statistical method used for modelling the relationship between a dependent variable and one or more independent variables. It's a method to find a straight line that best fits a set of data points. This line equation helps us understand the relationship between two things: one we want to predict (dependent variable) and one we use to make predictions (independent variable).

Some detailed explanation of what and how of linear regression:

1. Assumptions:

- **Linearity**: Assumes that there is a linear relationship between the independent and dependent variables.
- **Independence**: Assumes that the observations are independent of each other.
- **Homoscedasticity**: Assumes that the variance of the residuals (the differences between observed and predicted values) is constant across all levels of the independent variable(s).
- **Normality**: Assumes that the residuals follow a normal distribution.

2. Model Representation:

- The simple linear regression model can be represented as: $y=\beta 0+\beta 1 \cdot x+\epsilon$
    - y is the dependent variable.
    - x is the independent variable.
    - $\beta 0$ is the intercept (the value of y when $x=0$).
    - $\beta 1$ is the slope of the line (the change in y for a one-unit change in x).
    - ϵ is the error term (the difference between the observed and predicted values).

3. Objective:

- To minimize the sum of squared differences between the observed and predicted values, known as the **residuals**.

4. Ordinary Least Squares (OLS) Method:

- Linear regression uses a method called Ordinary Least Squares (OLS). It's like finding the line that minimizes the total distance from each point to the line.
- The OLS method calculates the coefficients ($\beta 0$ and $\beta 1$) that minimize the sum of squared residuals.

5. Parameter Estimation:

- **Intercept (β0)**: Represents the value of y when x=0. It is calculated as the mean of *y* minus the slope times the mean of x.
- **Slope (β1)**: Represents the change in y for a one-unit change in x. It is calculated as the covariance of x and y divided by the variance of x.
- Once the coefficients are estimated, the regression line equation is obtained.

6. Model Evaluation:

- **Coefficient of Determination (R2)**: Measures the proportion of the variance in the dependent variable that is predictable from the independent variable(s).
- **Residual Analysis**: Checking for patterns in the residuals to assess the model's adequacy.

Q2: Explain the Anscombe's quartet in detail.

Anscombe's quartet is a statistical phenomenon that demonstrates the limitations of relying solely on summary statistics (such as means, variances, and correlations) to understand datasets. It consists of four datasets, each containing 11 (x, y) pairs of data points. Despite the fact that these datasets have nearly identical summary statistics, they exhibit vastly different characteristics when visualised.

Here's a brief overview of each dataset in Anscombe's quartet:

1. Dataset I: This dataset forms a simple linear relationship between x and y, following the equation y=3x+2. It represents a straightforward case of linear regression.
2. Dataset II: Similar to Dataset I, but with one outlier. The relationship between x and y remains mostly linear, but the outlier significantly influences the regression line.
3. Dataset III: This dataset also appears to have a linear relationship between x and y, but it is not a simple case. The linear relationship breaks down if one were to consider just the summary statistics.
4. Dataset IV: This dataset does not follow a linear relationship. Instead, it demonstrates a curvilinear relationship, with x values clustered around the middle and y values varying more widely.

The significance of Anscombe's quartet lies in its demonstration of the importance of data visualisation. While summary statistics can provide useful information about a dataset, they may not reveal the full picture. Visualisation allows us to identify patterns, outliers, and relationships that summary statistics alone cannot capture.

Anscombe's quartet is often used to emphasize the importance of exploring data graphically before drawing conclusions. It serves as a cautionary tale against relying solely on numerical summaries, highlighting the potential for misleading interpretations when visual inspection is neglected.

By presenting four distinct datasets with similar summary statistics but vastly different underlying structures, Anscombe's quartet underscores the need for a comprehensive approach to data analysis that includes both quantitative and visual methods.

Q3. What is Pearson's R?

Pearson's r is a statistic used to measure the strength and direction of the linear relationship between two variables. It's commonly referred to as the Pearson correlation coefficient. This coefficient ranges from -1 to 1.

- If r is close to 1, it indicates a strong positive linear relationship, meaning that as one variable increases, the other variable tends to increase as well.
- If r is close to -1, it indicates a strong negative linear relationship, meaning that as one variable increases, the other variable tends to decrease.
- If r is close to 0, it suggests little to no linear relationship between the variables.

Pearson's r only measures linear relationships, so it might not capture other types of relationships (like curvilinear or non-linear ones). Additionally, it's sensitive to outliers and can be influenced by them.

Overall, Pearson's r is a widely used tool for assessing the association between two continuous variables, providing valuable insights into how they relate to each other in a linear fashion.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardised scaling?

Scaling is a process used in data preprocessing to transform the values of variables into a specific range or distribution. Scaling helps prevent features with larger scales from dominating those with smaller scales during model training.

Here's why scaling is performed to Equalize variable scales: Variables often have different units or ranges, making direct comparison difficult. Scaling puts them on a similar scale, making it easier to compare and interpret their relative importance.

There are two common types of scaling: normalized scaling and standardized scaling.

1. Normalized Scaling (Min-Max Scaling):
    - In normalized scaling, also known as min-max scaling, the values of variables are transformed to fall within a specific range, typically between 0 and 1.
    - The formula for min-max scaling is: $x' = (x - \min(x)) / (\max(x) - \min(x))$
    - This scaling method preserves the relative distances between data points and is useful when the distribution of the data is not necessarily Gaussian (normal).
2. Standardized Scaling (Z-score Scaling):
    - In standardized scaling, also known as z-score scaling, the values of variables are transformed to have a mean of 0 and a standard deviation of 1.
    - The formula for z-score scaling is: $x' = x - \text{mean}(x) / \text{std}(x)$

- This scaling method is particularly useful when the distribution of the data is approximately Gaussian, as it centers the data around 0 and ensures that it has a standard deviation of 1.

In summary, both normalized scaling and standardized scaling are important techniques in data preprocessing for machine learning. Normalized scaling is suitable when the range of values is known and bounded, while standardized scaling is useful when dealing with Gaussian-distributed data or when robustness against outliers is desired.

Q5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Yes, encountering infinite values for the Variance Inflation Factor (VIF) is not uncommon in certain situations. The VIF is used to detect multicollinearity among predictor variables in regression analysis. It measures how much the variance of an estimated regression coefficient is inflated due to collinearity with other predictors.

The formula for VIF for a predictor variable XiX_iXi is:

VIF(Xi) = 1/{1 - R_i^2}

where R_i^2 is the coefficient of determination (R-squared) obtained by regressing Xi against all the other predictor variables.

Now, if R_i^2 = 1, then the denominator in the formula becomes 1−1=01, resulting in a division by zero, which leads to an infinite VIF value.

This scenario happens when one predictor variable can be perfectly predicted by a linear combination of other predictor variables. In other words, there is perfect multicollinearity among the predictors. Perfect multicollinearity occurs when one or more predictor variables in a regression model are a perfect linear function of other predictor variables.

When perfect multicollinearity exists, it means that one or more predictor variables can be perfectly predicted from the others. This situation causes problems in regression analysis because it becomes impossible to separate the individual effects of the predictor variables on the response variable. Consequently, parameter estimates become unstable and uninterpretable.

In practical terms, encountering infinite VIF values serves as a warning sign that there are severe issues with multicollinearity in the dataset. Dealing with multicollinearity might involve removing redundant variables, transforming variables, to mitigate its effects.

Q6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A Q-Q plot, short for quantile-quantile plot, is a graphical tool used to assess whether a dataset follows a particular probability distribution, such as the normal distribution. It compares the quantiles of the dataset to the quantiles of a theoretical distribution, typically a normal distribution.

Here's how a Q-Q plot works:

1. Sorting the Data: First, the data points in the dataset are sorted in ascending order.
2. Calculating Theoretical Quantiles: Theoretical quantiles are calculated based on the chosen distribution (e.g., normal distribution) for the same number of observations as in the dataset.

3. Plotting: The ordered data points from the dataset are plotted against the corresponding theoretical quantiles. If the dataset follows the theoretical distribution closely, the points should fall approximately along a straight line.

The use and importance of a Q-Q plot in linear regression are as follows:

1. Assumption Checking: Linear regression models often assume that the residuals (the differences between the observed and predicted values) follow a normal distribution. A Q-Q plot of the residuals allows us to visually assess whether this assumption holds. If the residuals closely follow a straight line on the Q-Q plot, it suggests that the normality assumption is reasonable.

2. Detecting Outliers and Skewness: Deviations from a straight line on the Q-Q plot can indicate departures from normality, such as outliers or skewness in the data. Outliers may appear as points far away from the line, while skewness may manifest as curvature in the plot.

3. Model Validation: Q-Q plots can be used as part of model validation procedures to assess the adequacy of the linear regression model. A well-fitted model should have residuals that approximate a normal distribution, which would be reflected in a Q-Q plot showing points closely aligned with the theoretical line.

4. Decision Making: Q-Q plots provide a clear visual representation of how well the data conforms to the assumed distribution. This information can guide decisions about whether additional data transformations or model adjustments are necessary to improve the model's performance.

Overall, Q-Q plots are valuable diagnostic tools in linear regression analysis, helping analysts evaluate the validity of assumptions and the quality of the regression model. They provide insights into the distributional properties of the residuals, aiding in the interpretation and refinement of regression models.