Submitted by : Soumita Das

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 ultimate Multiple Linear Regression model contains many significant predictor variables, some of which are categorical in nature and have been transformed into dummy variables.
Categorical variable season_spring,mnth_sep,weekday_sat,weathersit-Light snow and Mist cloudy are significant in predicting the count of bike rentals

Below is the summary statistics of the optimised model with the significant input variables

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.827
Model:                            OLS   Adj. R-squared:                  0.824
Method:                 Least Squares   F-statistic:                     265.1
Date:                Wed, 11 Oct 2023   Prob (F-statistic):          5.29e-184
Time:                        22:17:14   Log-Likelihood:                 485.52
No. Observations:                 510   AIC:                            -951.0
Df Residuals:                     500   BIC:                            -908.7
Df Model:                           9
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
const                       0.2784      0.020     13.657      0.000       0.238       0.318
yr                          0.2369      0.008     28.103      0.000       0.220       0.253
workingday                  0.0529      0.011      4.628      0.000       0.030       0.075
temp                        0.3446      0.024     14.358      0.000       0.297       0.392
windspeed                  -0.1494      0.025     -5.885      0.000      -0.199      -0.100
season_spring              -0.1524      0.012    -12.317      0.000      -0.177      -0.128
mnth_sep                    0.0652      0.016      4.116      0.000       0.034       0.096
weekday_Sat                 0.0643      0.015      4.361      0.000       0.035       0.093
weathersit_Light Snow      -0.2842      0.025    -11.251      0.000      -0.334      -0.235
weathersit_Mist_Cloudy     -0.0810      0.009     -9.031      0.000      -0.099      -0.063
==============================================================================
Omnibus:                       59.521   Durbin-Watson:                   2.068
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              139.831
Skew:                          -0.621   Prob(JB):                     4.33e-31
Kurtosis:                       5.245   Cond. No.                         11.8
==============================================================================
```

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

Answer: Using ‘drop_first=True’ when creating dummy variables is crucial because it prevents the issue of multicollinearity in regression analysis.
Multicollinearity happens when two or more dummy variables are strongly related, which can cause unstable and untrustworthy regression models.
This eliminates redundancy and the associated multicollinearity problem.

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

Answer: 'atemp' has the highest correlation with target variable 'cnt' which is followed by 'temp' variable.
As per the correlation heatmap, correlation coefficient between atemp and cnt is 0.631. And correlation coefficient between temp and cnt is 0.627.

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

Answer: To validate the assumptions of the model and for ensuring the reliability for inference, have validated the below assumptions as follows:

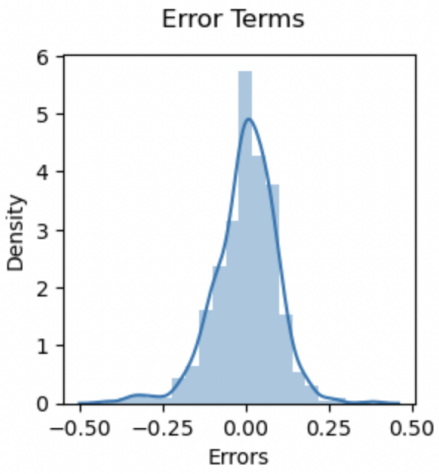
Residual Analysis:

Have checked the error terms if it is normally distributed (which is one of the major assumptions of linear regression). I have plotted the histogram of the error terms as mentioned below:

```
# Check the distribution of the error terms

fig = plt.figure(figsize=(3,3))
sns.distplot((y_train - y_train_pred), bins = 20)
fig.suptitle('Error Terms')
plt.xlabel('Errors')
```

```
Text(0.5, 0, 'Errors')
```



The residuals are following the normally distribution with a mean 0. All good!

Linear relationship between predictor variables and target variable:

This is happening because all the predictor variables are statistically significant (p-values are less than 0.05).

R-Squared value on training set is **0.827** and adjusted R-Squared value on training set is **0.824**. This means that most of the variance in data is being explained by these predictor variables.

Error terms are independent of each other:
The predictor variables are independent of each other. Multicollinearity issue is not there because the VIF (Variance Inflation Factor) for all predictor variables are below 5.
'Temp' has a VIF value just over 5 but it was not dropped because it has a strong correlation with the target variable as observed during EDA.

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 : Top 3 features significantly contributing towards demand of shared bikes are:
1) temp (coef: 0.2784)
2) yr (coef: 0.2369)
3) mnth_sep (coef: 0.0652)

General Subjective Questions

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

Answer: Linear Regression is a straightforward and widely used algorithm for predicting a numeric value (the "target") based on one or more input features (the "predictors"). It's like drawing a straight line through a cloud of points on a graph to make predictions.

Here's how it works:

Start with Data: You begin with a dataset that contains your target variable (the thing you want to predict) and one or more predictor variables (the things you think influence the target).

Find the Line: The goal is to find the best-fitting line (a straight line in the case of simple linear regression) that comes closest to touching all the data points. This line represents the relationship between the predictors and the target. The equation of the line is something like $y = mx + b$, where y is the target, x is the predictor, m is the slope, and b is the intercept.

Learn the Parameters: In simple linear regression, you use a process that calculates the best values for m (the slope) and b (the intercept). In multiple linear regression (with more than one predictor), you calculate a separate value for each predictor.

Make Predictions: Once you've learned the best-fitting line, you can use it to make predictions. For a new set of predictor values, you plug them into the equation, and it gives you a prediction for the target.
Evaluate the Model: To check how well your line fits the data, you can calculate an error metric, like Mean Squared Error (MSE) or R-squared, which tells you how close your predictions are to the actual values. The goal is to minimize this error.

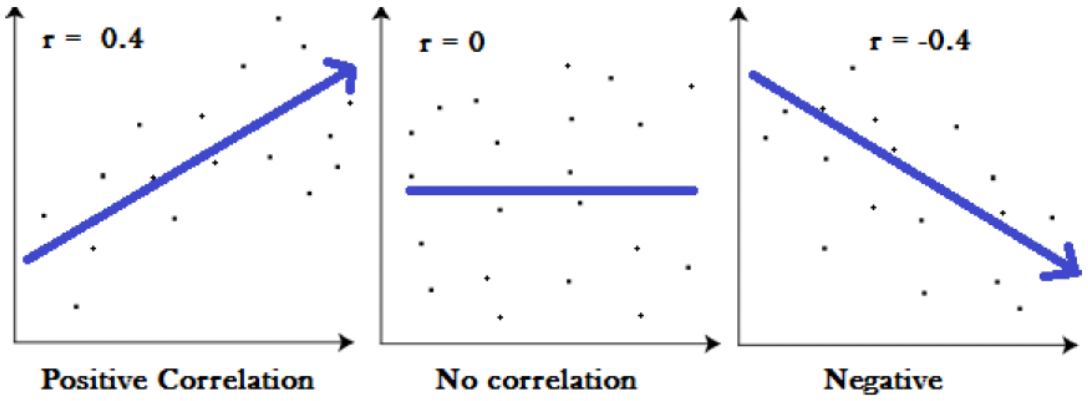
Linear regression is simple but powerful. The key idea is to find the best-fitting line that explains the relationship between variables and helps make accurate predictions.

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

Answer: Anscombe's Quartet is a small collection of four datasets created by the statistician Francis Anscombe. Each dataset has two variables (X and Y) and nearly identical statistical properties in terms of means, variances, and correlations. However, when you plot these datasets, you'll see that they have vastly different shapes and patterns. The purpose of the quartet is to emphasize the importance of data visualization in understanding the true nature of data and not relying solely on summary statistics. It illustrates that datasets with similar statistics can have distinct visual patterns, highlighting the need for exploratory data analysis and data visualization in statistical analysis.

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

Answer: Pearson's R, or the Pearson correlation coefficient, is a statistical measure used to quantify the strength and direction of the linear relationship between two continuous variables. It ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no linear correlation. It's a common tool for assessing the degree of association between two variables, such as the relationship between temperature and ice cream sales.



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer:

Scaling is a process in data preprocessing used to adjust the range of values in a dataset. It's done to make the data more suitable for machine learning models.

Why Scaling is Performed:

- Consistent Comparison: Scaling ensures that different features (variables) are on a similar scale. This allows for a fair comparison between them, as some machine learning algorithms are sensitive to the magnitude of variables.
- Faster Convergence: Scaling can help algorithms converge faster during training.
- Improved Model Performance: It can lead to better model performance, especially for methods that rely on distances or magnitudes of features.

Normalized Scaling vs. Standardized Scaling:

Normalized Scaling (Min-Max Scaling): It scales data to a specific range, often between 0 and 1. It's suitable when you know the minimum and maximum possible values for your data. It preserves the relative differences between values but not the mean or variance.

The general formula for normalization is given as:

Here, max(x) and min(x) are the maximum and the minimum values of the feature respectively.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Standardized Scaling (Z-Score Scaling): It transforms data to have a mean of 0 and a standard deviation of 1. It's used when the distribution of data is not known or is not assumed to be normal. It preserves the mean and variance of the data while making it centered at 0.
The general formula for Standardization is given as:
Here, σ is the standard deviation of the feature vector, and $\bar{x}$ is the average of the feature vector.

$$x' = \frac{x - \bar{x}}{\sigma}$$

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

Answer: A VIF (Variance Inflation Factor) can become infinite when perfect multicollinearity exists between predictor variables in a regression model. Perfect multicollinearity means that one predictor variable can be exactly predicted by a linear combination of other predictor variables, and this leads to unstable coefficient estimates. When perfect multicollinearity occurs, the VIF for the affected variable becomes infinite because it's impossible to quantify its unique contribution independently. This situation can arise due to data issues, such as duplicate variables or inappropriate model specification, and it typically requires addressing by removing or transforming variables to restore stability to the model.

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 used in statistics to compare the distribution of a dataset to a theoretical or expected distribution, often the normal distribution. It's a way to visually assess whether the data follows a particular distribution.

Use:
Distribution Comparison: In linear regression, you often make assumptions about the distribution of errors (residuals), like assuming they follow a normal distribution. Q-Q plots help you check if this assumption holds.
Identify Deviations: Q-Q plots display data points on a graph where the x-axis represents expected values from a theoretical distribution (e.g., normal) and the y-axis shows the actual data values. If the data follows the theoretical distribution, the points should form a straight line. Deviations from this line indicate departures from the assumed distribution.

Importance:
Assumption Checking: Linear regression relies on assumptions like normally distributed errors. If these assumptions are violated, the regression results may be unreliable. Q-Q plots allow you to detect deviations from the assumed distribution, helping you assess the validity of your model.

Data Transformation: If the Q-Q plot shows a systematic departure from the expected line, it can guide you in considering data transformations or alternative models to address the issue.
Model Improvements: Identifying deviations in the Q-Q plot may prompt you to improve your model by including additional variables, using different modeling techniques, or applying robust regression methods.