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)

    - Of all the categorical features, year has the largest effect on the independent variable. It has a moderate to high positive Pearson's r correlation factor of 0.54 indicating that as the year increases the dependent variable also increases.
    - Season does appear to impact the dependent variable with the values being lower in spring and higher in winter.
    - Month appears to have a smaller impact on the independent variable with January having the strongest negative impact on the dependent variable.
    - holiday, weekday, workingday and weathersit all have small impacts on the dependent variable.

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

    The drop_first parameter being set to True is important, without it the output of the pandas dummy method would inherently exhibit multicollinearity. This is because sum of all the dummy features will always be 1 if a column is not dropped. This will enable the calculation any dummy feature if the other features are known.

    if you have a categorical variable with three possible values, for example cat, dog and mouse, and you create a set of dummy variables (without dropping a column), you will have three columns (cat, dog and mouse). For each row only one will be populated with 1 and the others with 0. In this example, if cat == 0 and dog == 0 then you know mouse == 1. We also know that if cat == 1 then mouse == 0 and dog == 0.

    This can cause problems interpreting multiple linear regression coefficients as results may vary depending on which feature the model assigns the importance.

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

    Looking at the initial pair-plot, the temp and atemp features has the highest correlation (0.63) with the target variable.

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

    I attempted to validate each assumption in the following way:

    1) There is a linear relationship between each predictor variable and the response

    I created and reviewed a pair plot of the chosen features and looked for indications of a linear relationship with the target variable. I checked the Pearson's r statistic for each feature and ensured there was a correlation, and finally by checking the results of the t-test for each feature in the OLS model summary.

The pair plot did indicate some liner relationships between the features and the target variable (cnt), but it was not obvious.

The Pearson's r correlation coefficients provided a clearer indication of a linear relationship.

In the final model summary it was observed that all features have P>|t| < 0.05. This is sufficient to state that a relationship between the features and the target variable has been observed.

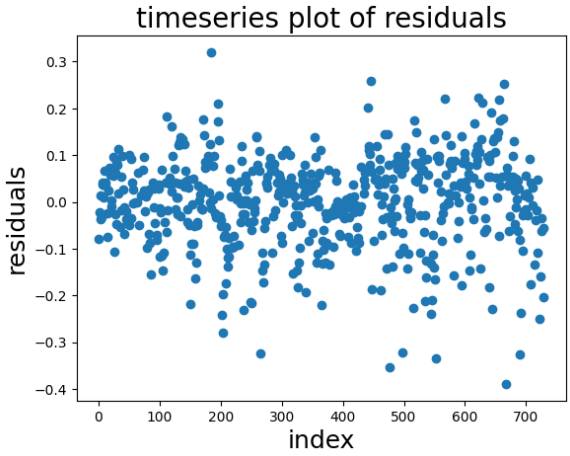
2) No Multicollinearity

This was confirmed by checking that the VIF scores of all features are less than 5. This ensures the R2-score for each feature is less than 0.80. So, no more than 80% of the variance of any feature can be explained using the other features in the dataset.

The VIF was calculated and was < 5 for all features apart from the constant. I don't think this is an issue as the constant has no variance. I think the high VIF may be due to the holiday field also having very little variance which is being mistaken for a correlation.

3) Observations are independent

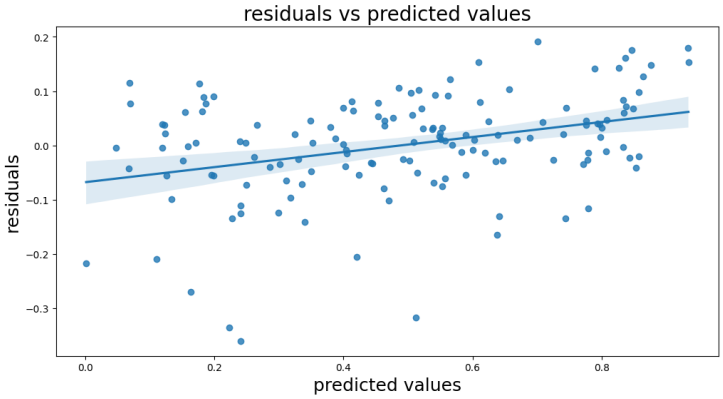
To do this I created a residual time series plot and checked to see if there was any visual indication of a relationship. I then checked the results of the Durbin-Watson test conducted during model fitting.



The plot created did not show any obvious relationship. This was confirmed with a Durbin-Watson test. In a DW test any value above 1.5 and below 2.5 is deemed as ok. The results of the DW test performed as part of the model fitting can be seen in the model summary. In this case the value was 1.975 and therefore fine.

4) Homoscedasticity

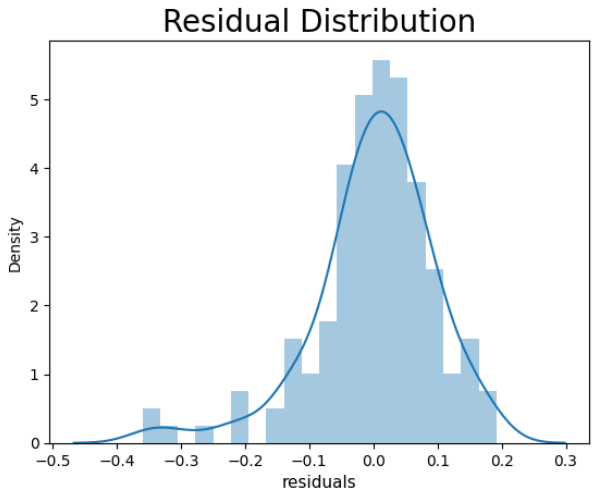
To test this, I created a scatter plot of the residuals against the predicted values and checked visually to see if a random distribution was observed.



The plot appears to show that the residuals have a mean of approximately 0 but a slight positive correlation (indicated by the regression line).

5) Multivariate normality

I checked this by plotting the distribution of the residuals and checking if it looked like a normal distribution.



The distribution plot showed that the errors are positively skewed, and the distribution is leptokurtic. This appears to indicate that the model is being impacted by outliers, but all outliers were removed during training. I would need more time to research on how to correct this.

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

Looking at the summary of the final model, the three features that have the largest absolute coefficient and therefore the largest impact on the target variable are:

1) Temperature (0.5715)

2) Year (0.2448)
3) Humidity (-0.2166)

General Subjective Questions

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

The linear regression algorithm is a machine learning algorithm that leverages a linear relationship between one independent variable (x) and one target variable (y) to identify an optimal linear equation that approximates the relationship between the two variables based on previous observations. The equation is of the form:

$$y = \beta_0 + \beta_1 x$$

In the equation β_0 and β_1 are model parameters that once calculated can be used to predict values of y when given an associated value of x.

An iterative algorithm called gradient descent is used to identify the optimal parameters for the equation.

Gradient descent is used to minimize a cost function. For linear regression the cost function is:

$$\frac{1}{2N} * \sum_{i=1}^{N} (Predicted\ y_i - Actual\ y_i)^2$$

Where N is the number of observations in the dataset.

By substituting predicted y_i with the linear equation we get:

$$\frac{1}{2N} * \sum_{i=1}^{N} (\beta_0 + \beta_1 x_i - Actual\ y_i)^2$$

To calculate the optimal values of each β the algorithm runs an iterative process. It starts with random values for β and calculates a new value using the following function:

$$\beta_{new} = \beta_{current} - \alpha(\frac{\delta L}{\delta \beta_{current}})$$

Where:

- α is a small positive step size value.
- $\delta L / \beta_{current}$ is the slope at the point of $\beta_{current}$ on a plot of the loss value against possible β values.

If the random starting point is too far left, the slope of the line is negative. As α is positive the value of the slope multiplied by α is also negative. This is then taken away from $\beta_{current.}$ Taking a negative number away from $\beta_{current}$ will see β_{new} increase, and therefore get closer to the minimum value. The lower the value of the slope the smaller the step that is taken towards

the optimal value. If the original value is too high, the slope will be positive so β_{new} will be smaller than $\beta_{current}$ and therefore still closer to the minimum value.

This process is repeated until $\beta_{new} = \beta_{current}$ or a predefined number of iterations has been completed.

The optimal values calculated for β_0 and β_1 can then be used to define the optimal linear equation.

$$y = \beta_0 + \beta_1 x$$

This can easily be scaled up for models with multiple features. In these instances, the linear equation becomes:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_n x_n$$

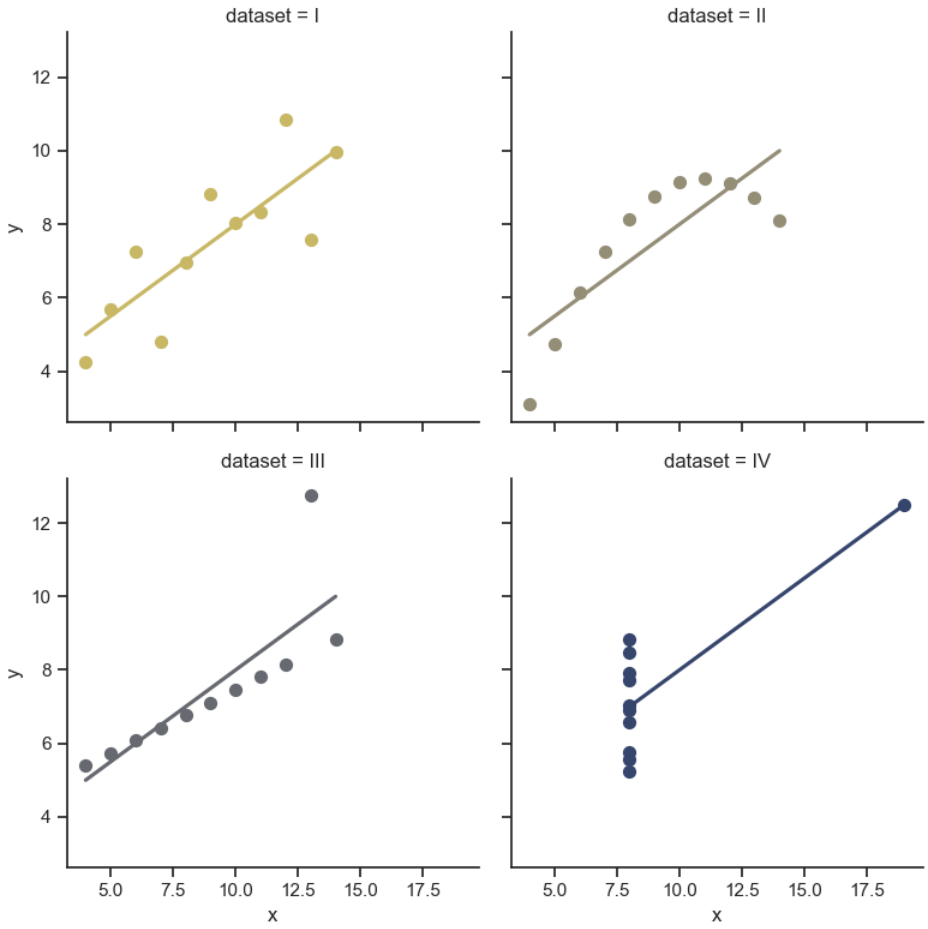
Where x_1 to x_n are the independent features and β_1 to β_n are the model parameters

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

Ansombe's quartet is a set of four datasets carefully curated to appear identical when comparing common statistical properties such as the means and standard deviations of the columns. If you perform linear regression the coefficient calculated for x is also very similar.

Property	Dataset 1	Dataset 2	Dataset 3	Dataset 4
count	11	11	11	11
X Mean	9	9	9	9
Y Mean	7.500909	7.500909	7.500000	7.500909
X standard deviation	3.316625	3.316625	3.316625	3.316625
Y standard deviation	2.031568	2.031657	2.030424	2.030579
Linear regression X coefficient	0.796803	0.796793	0.796673	0.796783

If you visualize these datasets in simple scatter plots the differences become obvious.



Plot created using code from [Anscombe's quartet — seaborn 0.12.2 documentation (pydata.org)]

The quartet illustrates the importance of simple visualisations

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

Pearson's R is a commonly used measure of linear correlation between two sets of data. The values of R always lie between 1 and -1 with:

- R = 1 signifying that there is a perfect positive correlation between the two sets of data. As one increases the other also increases in a proportional way
- R = 0 signifies no correlation between the two sets of data
- R= -1 signifying that there is a perfect negative correlation between the two sets of data. As one increases the other also decreases in a proportional way

Pearson's R is the default method for calculating correlation in the pandas corr() method. Examples of its use include spotting correlations between independent variables and the target and spotting and investigating multicollinearity between independent variables.

Pearson's R can be calculated for sets of data x and y using the following formula:

$$r_{xy} = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y}$$

Where:

$\sigma_x = standard\ deviation\ of\ x$
$\sigma_y = standard\ deviation\ of\ y$
$\mathrm{cov}(x, y) = the\ covariance\ of\ the\ x\ and\ y\ sets\ of\ data$

cov(x,y) can be calculated as follows:

$$\mathrm{cov}(x, y) = \sum_{i=1}^{N} \frac{(x_i - \mu_x) * (y_i - \mu_y)}{N - 1}$$

Where:

$N = The\ number\ of\ values\ in\ the\ sets\ of\ data$
$x_i = The\ value\ of\ x\ in\ row\ i$
$y_i = The\ value\ of\ y\ in\ row\ i$
$\mu_x = Mean\ of\ x$
$\mu_y = Mean\ of\ y$

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

Scaling is the process of taking the values of independent variables and sometimes the target variable and adjusting them to be in a consistent range. This is done to enable the coefficients calculated during multiple linear regression to be directly compared with each other to assess the significance and it also improves model fitting by helping gradient of descent find the optimal model parameters in less steps.

During normalization the values of the features or scaled to sit between 0 and 1. This is also called min-max scaling. This is achieved using the minimum and maximum values of the features in the following way:

$$x_{\mathrm{new}} = (x - x_{min})/(x_{max} - x_{min})$$

In standardization the values are transformed into a scale where the mean is zero and the standard deviation is one. This is done using the following formula:

$$x_{new} = \frac{x - \mu}{\sigma}$$

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

If a calculated value of VIF is infinite is means that all of the variance in the associated variable can be explained by other features in the model.

If we look at the formula for VIF

$$VIF_i = \frac{1}{1 - R_i^2}$$

Any value of R_i^2 >= 0 and R_i^2 < 1 will result in a real number value for VIF.

If R_i^2 == 1 then we will have:

$$VIF_i = \frac{1}{1 - 1}$$

$$VIF_i = \frac{1}{0}$$

1/0 is undefined but in this instance it's represented as INF.

When R_i^2 == 1, all of the variance in x_i can be explained by the other features

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

A Q-Q plot is a visualisation that illustrates how similar two sets of data are distributed. It works by plotting the quantile from one set of data against the same quantile from the other set of data. If the two distributions are equivalent the points on the plot should fall on a 45-degree line that runs through the origin (x=y line)

Q-Q plots are important for linear regression as they can be used to validate the assumption of Multivariate Normality by comparing the distributions of the residuals against the normal distribution.

If the distributions are not equivalent, the plot can give a indication of the nature of the differences in the distributions. This information could possibly be used to identify transforms that could be used to restore Multivariate Normality.