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?

    - **Working day** and **non-holiday** indicates that customers prefer using Bike sharing service on these days over holidays or weekends.
    - **Best** (Clear) weather situation is preferred over other weather situation, so target variable is dependent on '**weathersit'** categorical variable
    - **Fall** season shows more booking than any other season, however there's is not too much difference on average in the dataset over two years.
    - It can be observed that there is a gradual increase in booking over the years i.e. 2019 booking is more than 2018.

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

    - When using drop_first = True during dummy variable creation this code syntax drops 1 category from the categorical variable.
        - Eg. If there are 4 categories in a categorical variable then the above code will reduce it to 3.
    - This step is important in dummy variable creation because it avoids multicollinearity problem and also it reduces redundancy as the actual column doesn't lose information while doing this.

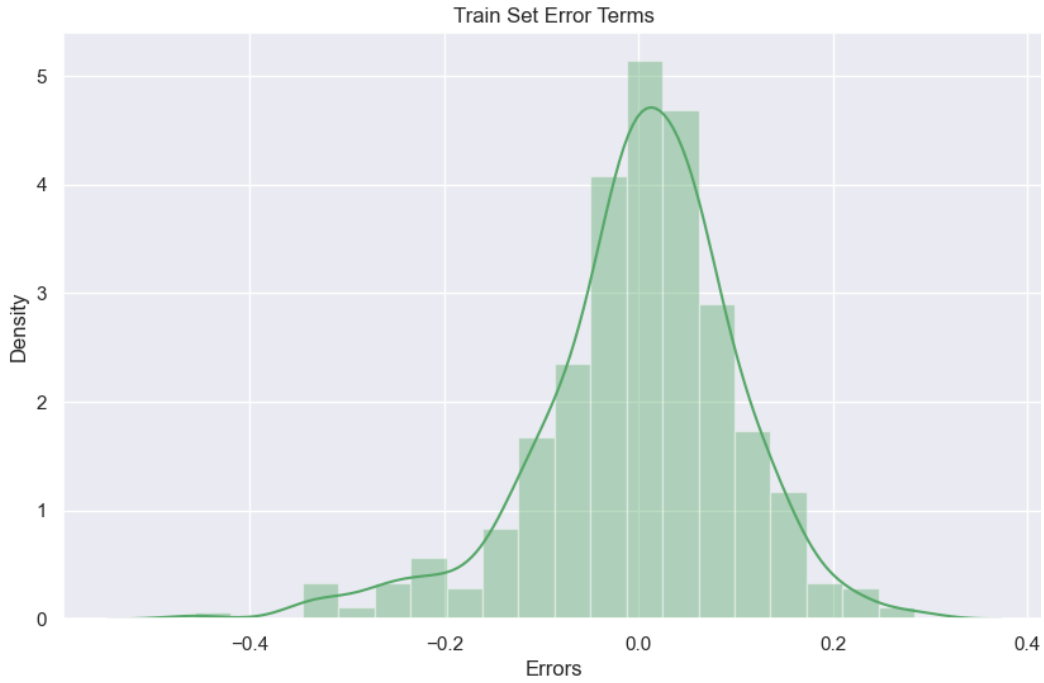
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

    - **Temp** variable shows highest correlation with the target variable however atemp is also close enough

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

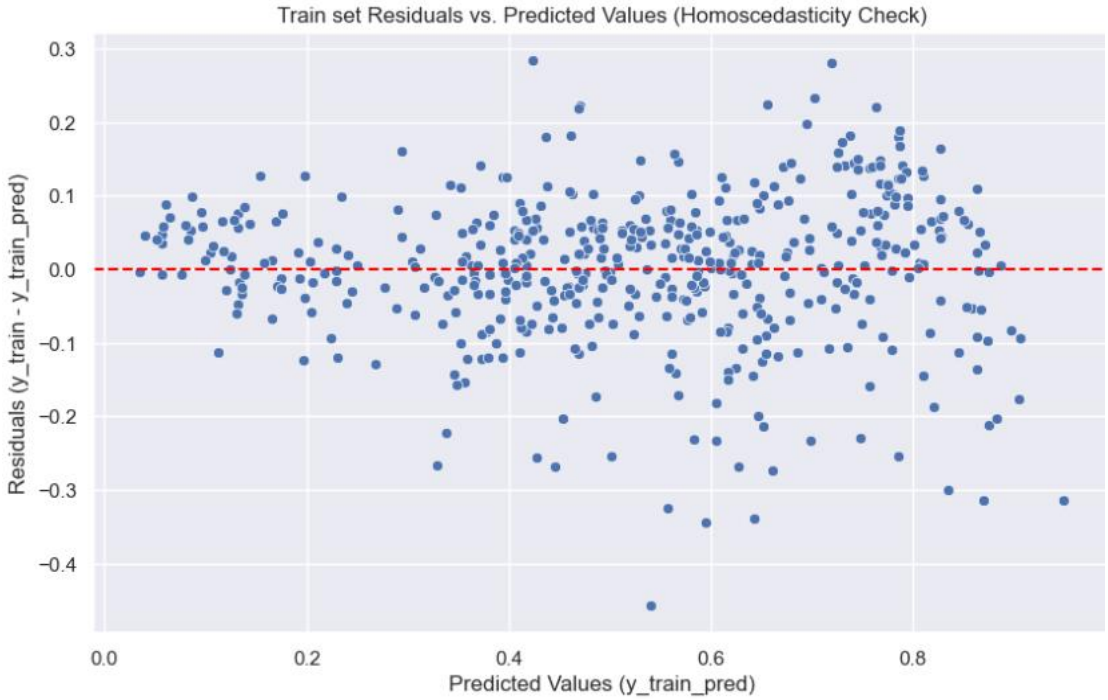
    - **Normality:**

        - Mean of the residual was -1.432256659718292e-15 which is approximately equal to zero suggesting that it follows a normal distribution



- **Homoscedasticity**:

  - Checked the residuals by plotting residuals vs fitted values as shown below. The spread of residual values is constant across all of the fitted values



- **Linearity:**

  - Above plot shows that the residuals are randomly scattered around zero line without any specific pattern, it indicates that the relationship between independent and target variable in linear in nature

- **Multicollinearity:**

  - Checked the variance inflation factor to determine if independent variables are correlated or not. Value less than 5 were selected as the threshold.

- **Independence:**

  - Model summary showed **Durbin-Watson** value as **2.196** which is very close to 2 which indicates that there is no significant autocorrelation in the residuals values which fulfils the independence assumption for the linear regression model.

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

   - Top 3 features contributing significantly towards explaining demand of the shared bikes are:
     - temp (Temperature)
     - yr (Year)
     - hum (Humidity)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

- Linear regression is a fundamental algorithm in statistics and machine learning used to model the relationship between a dependent variable and one or more independent variables

  - **Purpose:**

    Linear Regression aims to find the best fitting straight line that describes the relationship between the dependent variables and the target variable.

  - **Types of Linear Regression:**
    There are two types of Linear Regression 1. Simple Linear Regression – consisting of only 1 independent variable 2. Multiple Linear Regression – consisting of 2 or more independent variable

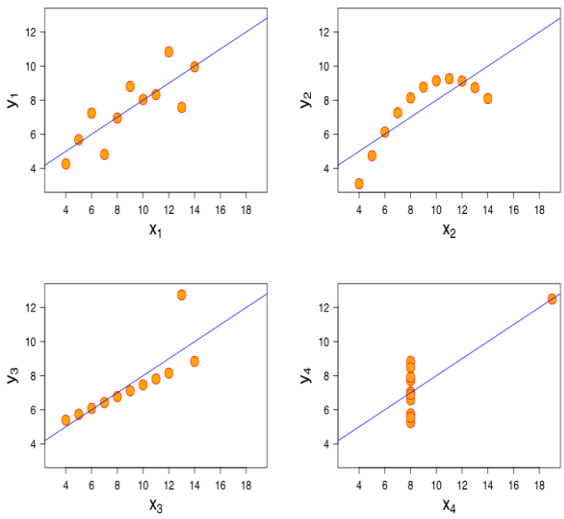
  - **Linear Regression Model:**
    Simple linear regression equation, y = b0 + b1x1 + e
    Multiple linear Regression equation, y = b0 + b1x1 + b2x2 + b3x3 ….+ bnxn + e
    Where:
    - y = target variable
    - x = independent variable
    - x1, x2, x3, xn represents independent variables up to n
    - b0 = intercept
    - b1, b2, b3, bn are the coefficients of independent variable which determines the importance of variable
    - e = error term

  - **Assumptions of Linear Regression:**
    - Linearity – relationship between dependent and target variable is linear
    - Independence – Observations are independent of each other
    - Homoscedasticity – Constant variance of error terms
    - Normality – Errors are normally distributed

  - **Coefficients:**
    Coefficients are estimated using the Ordinary Least Square Method (OLS) which minimizes the differences between the observed value and the predicted value

  - **Model Evaluation:**
    R-squared, Adjusted R-squared, p-value, F-statistic is used for model evaluation

2. Explain the Anscombe's quartet in detail.



Anscombe's quartet

Dataset I		Dataset II		Dataset III		Dataset 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	7.71
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	7.81	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	9.13	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

Anscombe's quartet is a group of datasets that have the same mean, standard deviation, and regression line, but which are qualitatively different.

It is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties.

Important points about these datasets:

- Dataset I: Appears to show simple linear relationship which makes it suitable for linear regression

- Dataset II: It shows no linear relationship and linear regression is not possible for this dataset

- Dataset III: It contains an outlier that affects the regression line which shows that how outliers influence the analysis

- Dataset III: It shows a vertical type line with one outlier at extreme right end.

3. What is Pearson's R?

- Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear correlation between two variables. It quantifies the strength and direction of the relationship between the variables.
- The value of Pearson's R ranges from -1 to 1 where **+1** indicates a positive linear relationship, **-1** indicates negative linear relationship and **0** indicates no linear relationship

  The formula for Pearson's R is:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

  Where:
  - x and y are two variables
  - $\bar{x}$ is the mean of x
  - $\bar{y}$ is the mean of y
  - xi is individual observation of x
  - yi is the individual observation of y
  - r is the Pearson correlation coefficient

- For the Pearson correlation coefficient, it is assumed that both x y variables are measured on a continuous scale and each variable is normally distributed

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

- Scaling is used for preprocessing the in machine learning to adjust the range of feature values in the dataset. This step ensures that all features contribute equally to the model when they have different units
- Scaling is performed to
  - Improve the model performance (e.g. Gradient descent works faster)

- It increases accuracy and reduces bias terms when features are scaled in a similar scale
- Models can be easier to interpret if different features are on the same scale

- **Normalization:** It scales the data to a definite range i.e. between **0, 1** and its formula is

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

It can be used when we want all the features in a similar scale between 0 and 1

- **Standardization:** It basically transforms the data in such a way that the mean of the data becomes 0 with a standard deviation of 1

$$X_{std} = \frac{X - \mu}{\sigma}$$

To use this, it is assumed that the data is normally distributed

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

- It was observed that sometimes the value of VIF were infinite. This suggests that there are some variables which have multicollinearity problem.
- It simply means that the one or more independent variable in the regression model is highly collinear with other independent variables
- To solve this problem, we need to remove such collinear variable as it does not provide any unique information during model building process.

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

- A Q–Q plot (quantile–quantile plot) is a probability plot, used for comparing two probability distributions by plotting their quantiles against each other.
- **Skewness checking:**
  - Q-Q plots are also used to find the skewness of a distribution. When we plot theoretical quantiles on the x-axis and the sample quantiles whose distribution we want to know on the y-axis, then we see a very peculiar shape of a normally distributed Q-Q plot for skewness. If the bottom end of the Q-Q plot deviates from the straight line but the upper end does not, then we can clearly say that the distribution has a longer tail to its left.
  - When we see the upper end of the Q-Q plot deviate from a straight line while the lower follows one, then the curve has a longer tail to its right and it is right-skewed, also called positively skewed.
- **Normality Checking:**
  - If the two distributions that we are comparing are exactly equal, then the points on the Q-Q plot will perfectly lie on a straight-line y = x
  - We can visually check if the residuals on a Q-Q plot follows normal distribution or not, if it follows then we can say that the residuals are normally distributed.