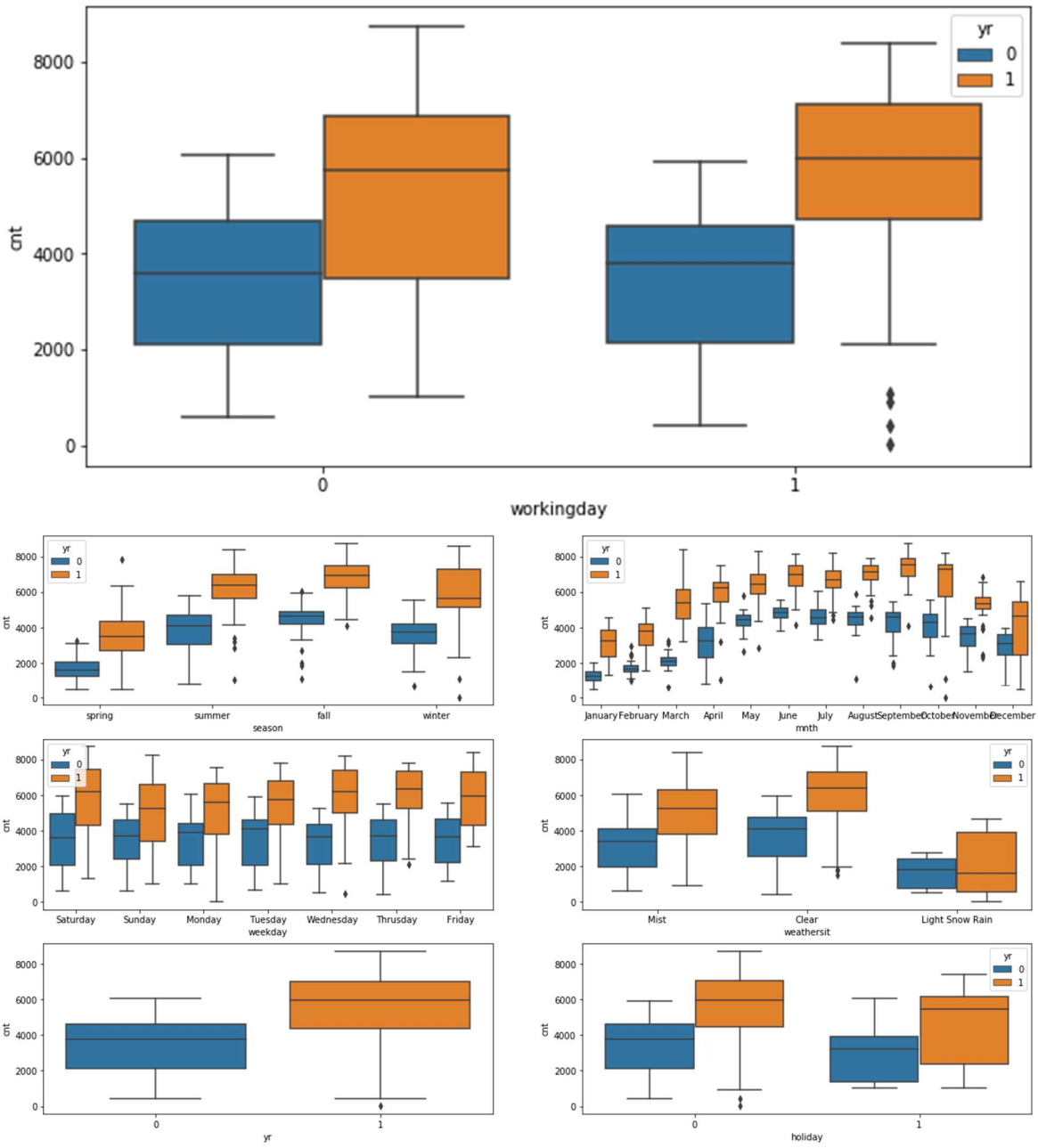
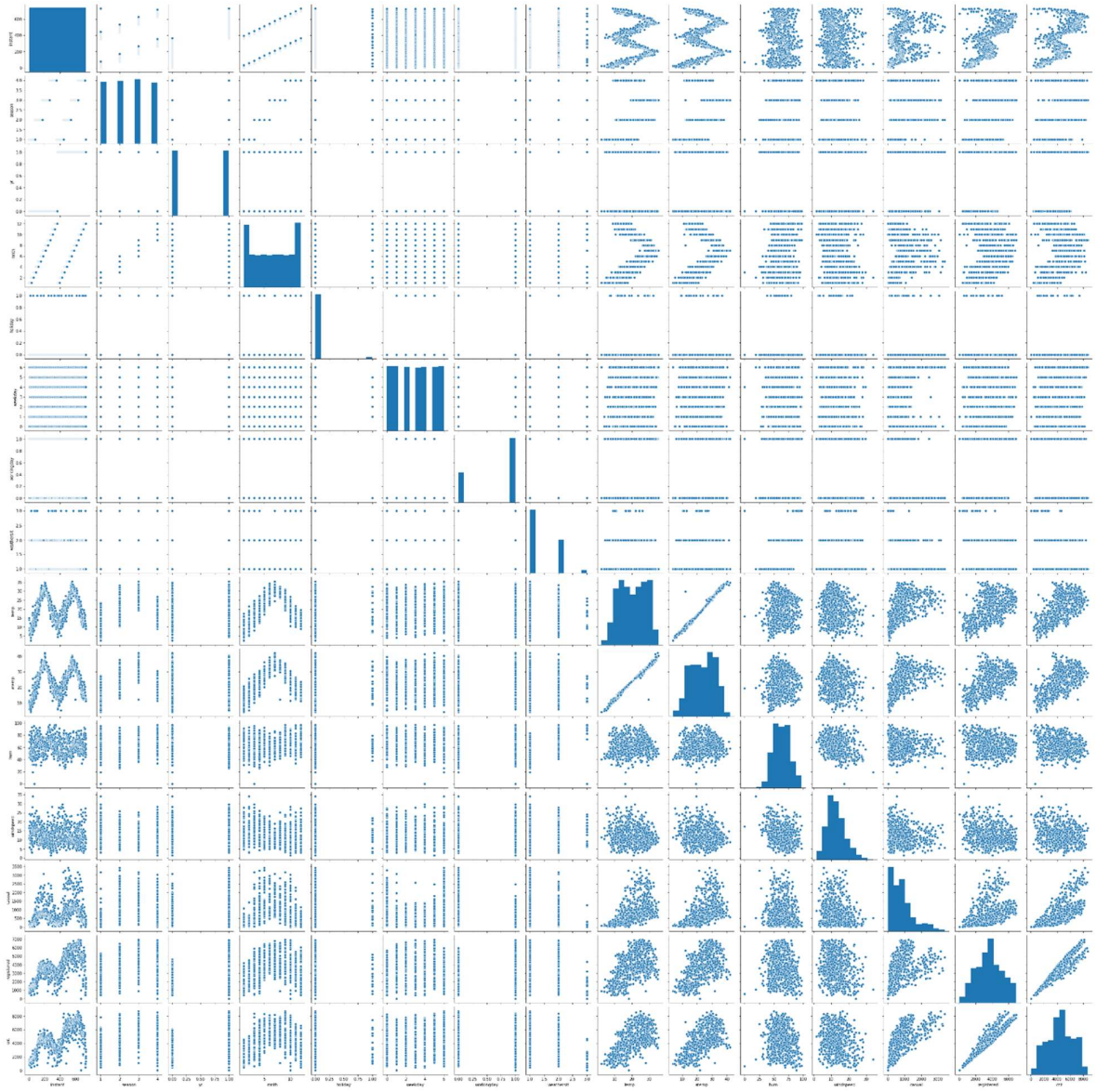
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
    i. Looking at all the variables we could say that sales/rental of bikes for year 2019 seems much higher as compared to 2018.
    ii. In 'mnth' (month) variable we can see that the sales/rental of bikes increases from month of March to September which is from 'summer' to 'fall' and this may be because these months are the beginning of vacations.
    iii. During Snow Rains there is drastic fall in the sales of bikes as compared of Mist and Clear weather for both years.
    iv. Also, during holidays, we can see that people buy that much cars because they are involved in spending holidays by roaming or partying with families, etc.
    v. While 'workingday' feature has not much impact on the count of sales of bikes.

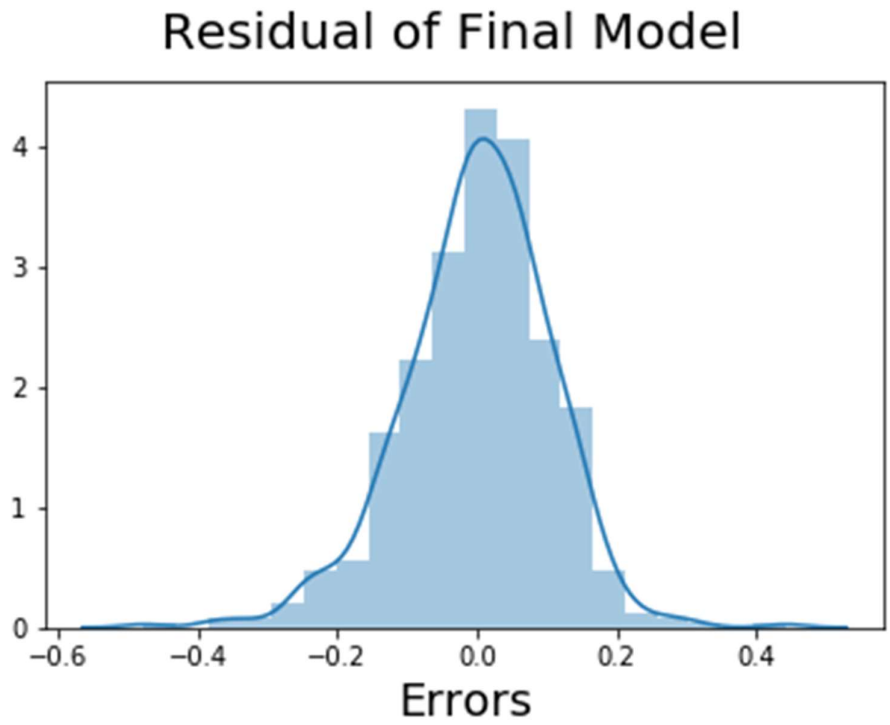


2. Why is it important to use **drop_first=True** during dummy variable creation?
    i. drop_first = True, allows us to drop the column to avoid redundancy.
    ii. Using drop_first = True in get_dummies() completely depends on the model, but if we don't drop the first column then your dummy variables will be correlated.
    iii. This may affect some models adversely and the effect is stronger when the cardinality is smaller.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
    i. On the basis of analysis, 'registered' variables seem to have highest correlation with target variable 'cnt'.



4. How did you validate the assumptions of Linear Regression after building the model on the training set?
    i. I had used VIF values and p-values to understand which variables had high correlation and were insignificant, and dropped those variables to avoid multicollinearity.
    ii. I used the Residual analysis to verify the Linear Regression assumptions after building the model on the training set where if the Residual/Error mean is zero this means the model is good to run with test set for validation.
    iii. Lastly the r2_score gives the final analysis.



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
    i. 'yr' variable which represent Year(2018 & 2019)
    ii. 'spring' variable
    iii. 'Light Snow Rain' variable

General Subjective Questions

1. Explain the linear regression algorithm in detail.
    a. Machine Learning has three types of Algorithms: Regression, Classification and Clustering and these algorithms are divided under two learning methods namely Supervised and Unsupervised.
    b. Regression Algorithm comes under Supervised learning method which means it has labels that is past information of the data set.
    c. Regression model has Simple Linear Regression (SLR) and Multiple Linear Regression (MLR) Algorithms under its heading.
    d. Both regression algorithms deal with continuous labels/values.
    e. SLR represents relationship between one independent variable (X-axis) and one dependent variable(Y-axis) whereas MLR represents relationship between two or more independent and a dependent variable.
    f. Linear Regression works on Straight Line formula to predict the Best Line Fit on the y_test and y_pred data (i.e. actual data and predicted data).
    g. Formula: y = $y = \beta 0 + \beta 1 . x$
        - where, β0 is slope
        - β1 is intercept
        - x and y are independent and dependent variables respectively

2. Explain the Anscombe's quartet in detail.
    a. Anscombe's quartet are the graphical representation of the four data sets that are almost identical with many similarities in their statistical properties.
    b. This demonstration explains that even though Statistics is powerful in describing general trends and aspects of data, it alone can't fully depict any data set.
    c. Though all four of these data sets have the same variance in x, variance in y, mean of x, mean of y and linear regression, we cannot accurately portray data in its native form.
    d. Anscombe's quartet explains the danger of outliers in data sets.

3. What is Pearson's R?
    a. Correlation Coefficient are used in Statistics to measure the relationship between two variables.
    b. Correlation Coefficient are of several types, but among those Pearson's Correlation or Pearson's R is a most frequently used correlation coefficient in Linear Regression.
    c. The formula of Pearson's R returns a value between -1 to 1, where
        - 1 indicates high positive correlation
        - -1 indicates high negative correlation
        - 0 indicates no relation i.e. neutral correlation

d. A correlation coefficient of 1 means that for every positive increase in one variable, there is positive increase of a fixed proportion in other.
e. A correlation coefficient of -1 means that for every positive increase in one variable, there is negative decrease of a fixed proportion in other.
f. A correlation coefficient of 0 means that for every increase or decrease in one variable, there is no positive or negative increase/decrease, which means they are not related to each other.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
    a. Scaling or Feature Scaling is a method used to normalize the range of independent variables or features of data.
    b. This method is performed during data preprocessing step and it is also known as data normalization in data processing.
    c. Feature scaling makes it easy to interpret the data and also helps in faster convergence for gradient descent method.
    d. Feature scaling has two methods to normalize the data:
        - Standardizing Scaling
        - Min-Max Scaling
    e. Standardization brings all the data into a standard normal distribution with mean as 0 and standard deviation as 1
        - Formula: $x = \frac{x - m \quad (x)}{sd(x)}$
    f. Min-Max scaling in the other hand brings the data in the range of 0 to 1
        - Formula: $x = \frac{x - \min(x)}{\max(x) - \min(x)}$
    g. Scaling is important because if we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients and this could lead to poor prediction rate by model.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
    a. VIF or Variance Inflation Factor helps to detect features with high multicollinearity.
    b. The higher the VIF value, the greater is the correlation of the variable with other variables.
    c. Features with value greater than 5 are regarded as variables with high multicollinearity and those need to be dropped (Sometimes features with value greater than 2 are also considered to be dropped due to high multicollinearity).
    d. If there is a perfect correlation, then VIF is infinite.
    e. An infinite VIF value indicates that corresponding variable may be expressed exactly by linear combination of the other variables (which show an infinite VIF as well).

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
    a. Q-Q or Quantile-Quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, Exponential or Uniform distribution.
    b. Also, it helps to determine if two data sets come from populations with a common distribution.
    c. Q-Q plot can be used with sample sizes also and many distributional aspects like shifts in location, shifts in scale, changes in symmetry and the presence of outliers can all be detected from this plot.
    d. It basically helps to determine whether two data sets come from populations with a common distribution.
    e. Identifies if two data sets have common location and scale.
    f. Identifies if two data sets have similar distributional shapes.
    g. Identifies if two data sets have similar tail behavior.
    h. If all points of quantiles lie on or close to straight line at an angle of 45 degree from X-axis then two data sets have similar distribution.
    i. If y-quantiles are lower than the x-quantiles than Y-values < X-values.
    j. If x-quantiles are lower than the y-quantiles than X-values < Y-values.
    k. If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis then two data sets have different distribution.