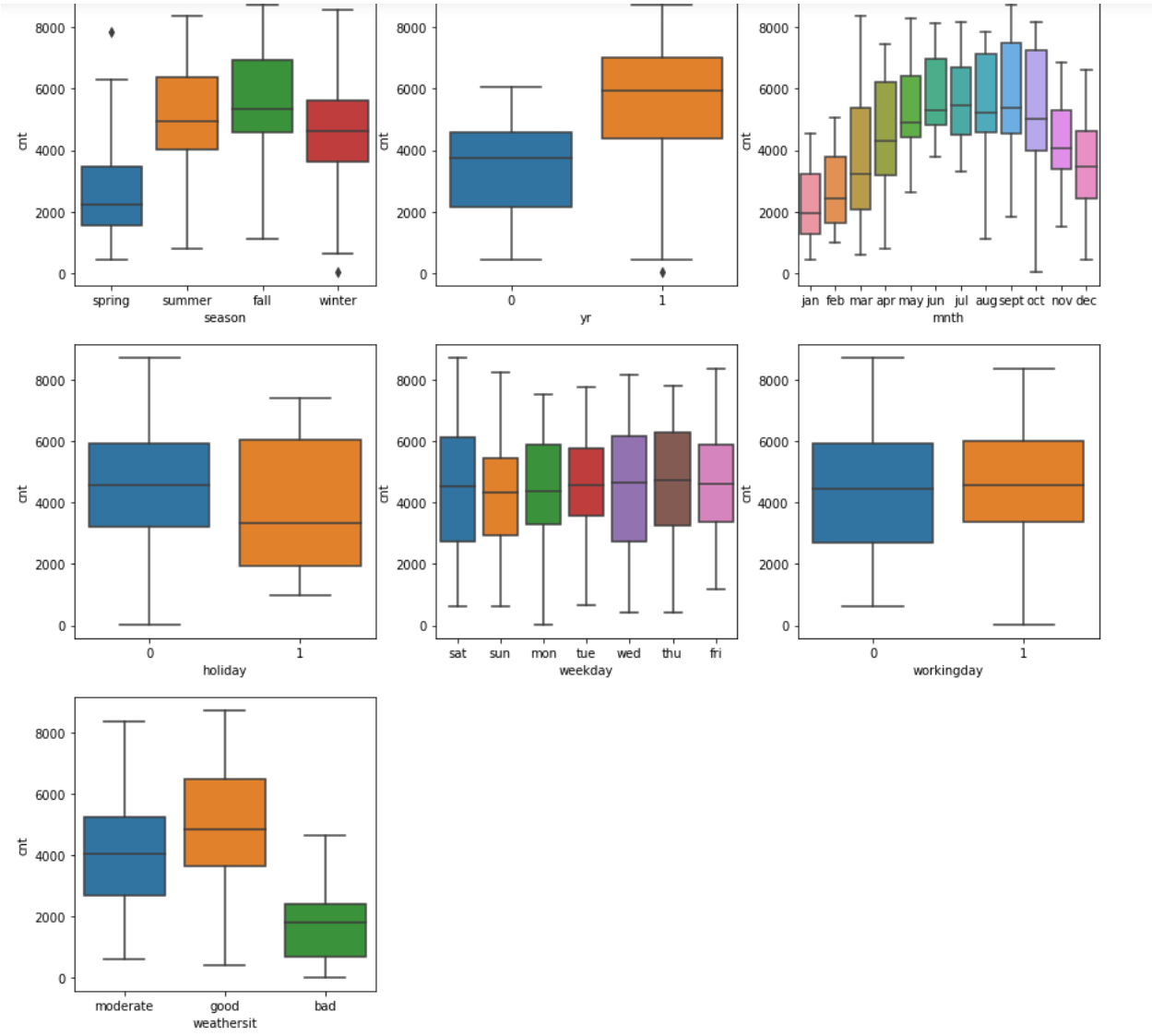
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Ans: There are categorical variables namely season, mnth, yr, weekday, working day and weathersit. These categorical variables have a major effect on the dependent variable 'cnt'. The below fig shows the correlation among the same.



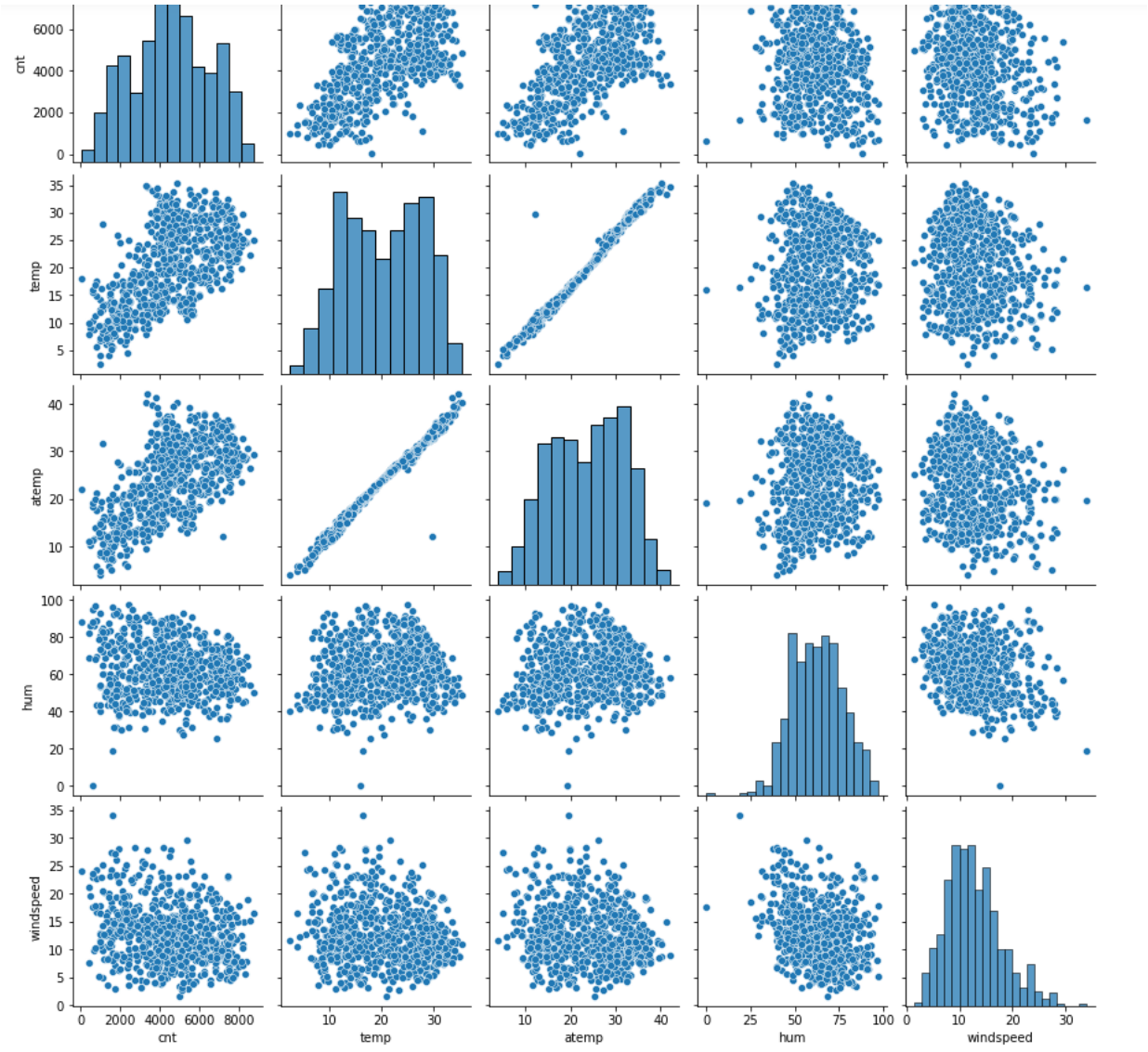
2. Why is it important to use drop_first=True during dummy variable creation?

Ans: The intention behind the dummy variable is that for a categorical variable with 'n' levels, you create 'n-1' new columns each indicating whether that level exists or not using a zero or one. Hence drop_first=True is used so that the resultant can match up n-1 levels. Hence it reduces the

correlation among the dummy variables. Eg: If there are 3 levels, the drop_first will drop the first column

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

Ans: From the pair plots we can say that temp and atemp has a highest correlation compared any other variables.



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

Ans: Linear Regression models are validated based on Linearity, No auto-correlation, Normality of errors, Homoscedasticity, Multicollinearity.

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

Ans: Temperature, Year, Season

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:

$$y=mx+b$$

Linear regression is a predictive modeling technique which tells us the relationship between the dependent (target variable) and independent variables (predictors). Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables. A regression line can be a Positive Linear Relationship or a Negative Linear Relationship. The goal of the linear regression algorithm is to get the best values for a0 and a1 to find the best fit line and the best fit line should have the least error. In Linear Regression, RFE or Mean Squared Error (MSE) or cost function is used, which helps to figure out the best possible values for a0 and a1, which provides the best fit line for the data points.

2. Explain the Anscombe’s quartet in detail.

Ans:

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, including mean, variance, correlation, and linear regression lines, but are visually distinct and demonstrate the importance of graphically exploring your data. The quartet was created by the statistician Francis Anscombe in 1973 to illustrate the limitations of relying solely on summary statistics without visualizing the data.

The four datasets in Anscombe's quartet are denoted as I, II, III, and IV. Each dataset consists of 11 (x, y) pairs. Here are the characteristics of the quartet:

Dataset I: A simple linear relationship. The relationship between x and y is well-described by a linear regression.

Dataset II: Similar to Dataset I but with one outlier that significantly influences the linear regression line. Removing the outlier would result in a completely different regression line.

Dataset III: A non-linear relationship. The data points follow a clear quadratic pattern.

Dataset IV: An extreme case where all variables are the same except for one point, which is an outlier. This dataset demonstrates the impact of leverage points on linear regression.

3. What is Pearson's R?

In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables.

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

Ans: Scaling means you're transforming your data so that it fits within a specific scale. It is one type of data pre-processing step where we will fit data in specific scale and speed up the calculations in an algorithm. Collected data contains features varying in magnitudes, units and range. If scaling is not performed than algorithm tends to weigh high values magnitudes and ignore other parameters which will result in incorrect modeling

1. In normalized scaling minimum and maximum value of features being used whereas in Standardize scaling mean and standard deviation is used for scaling
2. Normalized scaling is used when features are of different scales whereas standardized scaling is used to ensure zero mean and unit standard deviation.
3. Normalized scaling scales values between (0,1) or (-1,1) whereas standardized scaling is not having or is not bounded in a certain range.
4. Normalized scaling is affected by outliers whereas standardized scaling is not having any effect by outliers.
5. Normalized scaling is used when we don't know about the distribution whereas standardized scaling is used when distribution is normal.
6. Normalized scaling is called as scaling normalization whereas standardized scaling is called as Z Score Normalization.

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

Ans:

Formula for VIF is 1/(1-R2) . if its perfectly correlated with any other variable which correlation 1, then result of VIF will leads to infinite.

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

Q–Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data possibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.

QQ plot can also be used to determine whether or not two distributions are similar or not. If they are quite similar you can expect the QQ plot to be more linear. The linearity assumption can best be tested with scatter plots. Secondly, the linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot.

Importance of QQ Plot in Linear Regression :

In Linear Regression when we have a train and test dataset then we can create Q-Q plot by which we can confirm that both the data train and test data set are from the population with the same distribution or not.

Advantages:

- It can be used with sample size also
- 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

Q-Q plot use on two datasets to check

- If both datasets came from population with common distribution
- If both datasets have common location and common scale
- If both datasets have similar type of distribution shape
- If both datasets have tail behavior