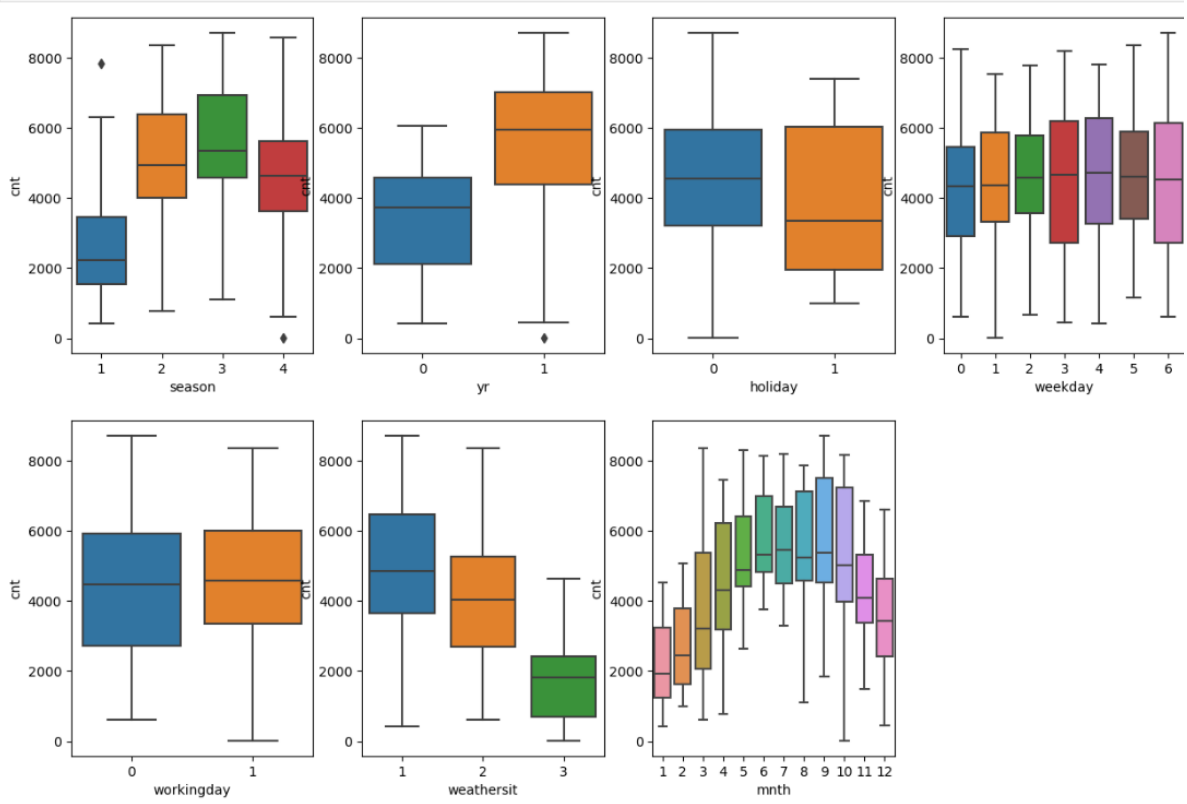
Assignment-based Subjective Questions

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

   Answer: There are categorical variables like season, year, month, weather situation and working day which have a major affect on bike counts 'cnt'. Detailed analysis below -
   season: Almost 32% of the bike booking were happening in season3 with a median of over 5000 booking (for the period of 2 years). This was followed by season2 & season4 with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable."

   - mnth: Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.
   - Yr: sales have increased drastically in 2019 compared to 2018 hence indicated that it is a good predictor.
   - weathersit: Almost 67% of the bike booking were happening during 'weathersit1 with a median of close to 5000 booking (for the period of 2 years). This was followed by weathersit2 with 30% of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.
   - holiday: Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.
   - weekday: weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.
   - workingday: Almost 69% of the bike booking were happening in 'workingday' with a median of close to 5000 booking (for the period of 2 years). This indicates, workingday can be a good predictor for the dependent variable

Below is the boxplot of categorical variables -



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

   Answer - If we have a categorical variable with n number of categories, when we create dummy variables we create n columns where each represents one category. Now, we we know the value of n-1 categories, we automatically know the value of the nth category. Therefore, they are highly correlated.

   This multi-collinearity can lead to problems like overfitting in the model because of redundant information. It can make the model's parameters' estimates to be undefined, making interpretation difficult.

   Hence by using "drop_first=True", we are removing the first level of the categorical variable, hence we end up with n-1 dummy variables, thus avoiding the dummy variable trap. This doesn't result in any loss of information because the dropped category can be represented as all zeros.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?** (1 mark)
   Answer : The temperature 'temp' and feeling temperature 'atemp' variables have highest correlation when compared to the rest with target variable bike count 'cnt'

4. **How did we validate the assumptions of Linear Regression after building the model on the training set?** (3 marks)
   Answer: After building a Linear Regression model on the training set, we can validate its assumptions using below techniques:

1. Linearity: We can create scatter plots of the predicted vs actual values to see if they follow a linear pattern. If the relationship is non-linear, we can consider transforming the input, the output, or both.
2. Independence: This assumption states that the residuals should be independent. Any trend or pattern in the residuals would suggest that there is some information that the model failed to capture.
3. Homoscedasticity: This assumption states that the residuals should be independent. Any trend or pattern in the residuals would suggest that there is some information that the model failed to capture.
4. Multivariate Normality: This assumes that the residuals are normally distributed. This can be checked using a Q-Q plot, where the residuals are plotted against a theoretical normal distribution in such a way that the points should form an approximate straight line
5. Lack of Multicollinearity: This assumption states that the predictor variables should not be highly correlated with each other. This can be checked using the Variance Inflation Factor (VIF), where a VIF of 1 indicates no correlation, and a VIF greater than 5 or 10 indicates high multicollinearity.

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 contributing significantly towards explaining the demand of the shared bikes are –

   1.With an increase in temperature the demand also increases, hence it should keep track of the weather conditions.

   2. Year – count of bikes increases with every hear

   3. Season – season significantly affects the count of bikes being rented.

   Look at the coefficients of these variables temp, year and season in the below model output

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.835
Model:                            OLS   Adj. R-squared:                  0.832
Method:                 Least Squares   F-statistic:                     253.7
Date:                Wed, 26 Jun 2024   Prob (F-statistic):          1.13e-188
Time:                        18:28:42   Log-Likelihood:                -4135.7
No. Observations:                 511   AIC:                             8293.
Df Residuals:                     500   BIC:                             8340.
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        668.8308    161.877      4.132      0.000     350.789     986.873
yr          2030.8265     71.448     28.424      0.000    1890.452    2171.201
workingday   492.2594     97.148      5.067      0.000     301.390     683.129
temp        4781.2601    171.745     27.839      0.000    4443.830    5118.690
windspeed  -1347.5665    217.998     -6.182      0.000   -1775.871    -919.262
season_2     769.9694     89.584      8.595      0.000     593.963     945.976
season_4    1145.8105     89.967     12.736      0.000     969.050    1322.571
weekday_6    587.6809    125.150      4.696      0.000     341.797     833.565
weathersit_2 -699.5312    76.114     -9.191      0.000    -849.074    -549.988
weathersit_3 -2501.6370   215.126    -11.629      0.000   -2924.300   -2078.974
mnth_9       844.1419    137.001      6.162      0.000     574.973    1113.311
==============================================================================
Omnibus:                       68.737   Durbin-Watson:                   2.081
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              152.172
Skew:                          -0.730   Prob(JB):                     9.04e-34
Kurtosis:                       5.239   Cond. No.                         11.6
```

General Subjective Questions

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

Answer - Linear Regression is one of the most fundamental and widely known Machine Learning Algorithms which people start with building their career in Machine Learning. It's a statistical method to find the relationship between one dependent variable (usually denoted as Y) and one or more independent variables (usually denoted as X).

Here are the details of how Linear Regression works:

1. **Model Building**: The algorithm uses the relationship Y = b0 + b1*X + e for simple linear regression and Y = b0 + b1*X1 + b2*X2 + ... + bn*Xn + e for multiple linear regression. Here, b0 is the y-intercept, b1 is the slope of the line (or coefficients of X in multiple regression), and e is the error term.
2. **Learning**: The algorithm calculates the best estimate for the regression coefficients b0 and b1 using the method of least squares. The least squares method minimizes the sum of the squared residuals (the differences between the actual and predicted values).
3. **Prediction**: Once the coefficients are estimated, the model can predict the response (Y) for a new set of predictor variables (X).
4. **Evaluation**: The model's performance is evaluated using metrics like R-squared, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), etc.

   It's important to note that the underlying assumptions of the Linear Regression model are that there is a linear relationship between the dependent and independent variables, the residuals are normally distributed and have constant variance (homoscedasticity), and there is no multicollinearity (high correlation between predictor variables).

Linear Regression is a simple yet powerful algorithm, widely used in various domains, from economics to machine learning, because of its interpretability and computational efficiency.

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

Answer - Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

Here's a basic summary of the four datasets:

1. The first dataset appears to be a simple linear relationship, corresponding to two variables correlated and following the assumption of normality.
2. The second dataset is not distributed normally; while an obvious relationship between the two variables can be observed, it's not linear, but curve shaped.
3. The third dataset is distributed linearly but with a different regression line, which is offset by the presence of an outlier – it introduces increased variability in the y variable (also known as heteroscedasticity).
4. The fourth dataset shows an example when one outlier is enough to produce a high correlation coefficient, even though the relationship between the two variables is not linear.

Despite the differences in these datasets:

- All four datasets have the same mean of the x and y variables.
- The variance of the x and y variables is identical in the four datasets.
- The correlation between x and y is the same in all four cases.
- The regression line (the line that best fits the data) is the same for all four datasets.

Anscombe's quartet is a demonstration of why it is essential to not only rely on statistical properties when analyzing data, but also visualize the data to understand the context and see if any patterns or anomalies exist.

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

Answer - Pearson's R, also known as Pearson's Correlation Coefficient, is a statistical measure that calculates the strength and direction of the linear relationship between two variables. The values of Pearson's R range from -1 to +1.

Here's what the values mean:

- A coefficient of +1 indicates a perfect positive linear relationship between the variables. That is, as one variable increases, the other variable also increases.
- A coefficient of -1 indicates a perfect negative linear relationship between the variables. That is, as one variable increases, the other variable decreases.
- A coefficient of 0 indicates no linear relationship between the variables.

The formula for calculating Pearson's R is the covariance of the two variables divided by the product of their standard deviations.

It's an essential tool in the fields of mathematics and statistics and widely used in data science and machine learning to understand the linear relationship between input features and to deal with multicollinearity in datasets.

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

Answer - Scaling is a preprocessing technique used in machine learning and data processing. It is the process of transforming the data into a standard scale, allowing for comparison between variables that may have different units or ranges.

Scaling is performed for several reasons:

1. Machine learning algorithms often perform better or converge faster when features are on a relatively similar scale.
2. It prevents certain features from dominating others due to differences in units or range.
3. It deals with the issue of outliers as scaling reduces the impact of outliers on the model.
4. Some algorithms, like K-Nearest Neighbors (KNN) and Gradient Descent, require features to be on the same scale for the model to perform correctly.

There are two common types of scaling: Normalization and Standardization.

Normalization: This method rescales the features to a range of [0, 1]. This might be useful in some cases where all parameters need to have exactly the same positive scale. However, the outliers from the data are lost. The formula used for normalization is (X - X.min) / (X.max - X.min)

Standardization: This method transforms the data to have zero mean and a variance of 1. This standardization assumes that wer data follows a Gaussian distribution (bell curve). This doesn't necessarily bound values to a specific range as normalization does. The formula used for standardization is (X - mean(X)) / std_dev(X)

The choice of scaling method depends on the algorithm used and the specific distribution of the data.

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

Answer - The Variance Inflation Factor (VIF) is a measure of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable. This ratio is calculated for each independent variable. A high VIF indicates that the associated independent variable is highly collinear with the other variables in the model.

VIF becomes infinite when the denominator in the VIF formula becomes zero. This happens when the independent variable can be expressed perfectly by a linear combination of other variables, meaning extreme multicollinearity exists. In other words, one independent variable is a perfect predictor of another independent variable.

An infinite VIF is a strong indication that we need to revise wer model. This could involve removing variables that are dependent on others or combining variables that are measuring the same underlying concept.

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 (Quantile-Quantile) plot is a graphical tool to help us assess if a dataset follows a given theoretical distribution. It is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points

forming a line that's roughly straight.

In the context of linear regression, a Q-Q plot is used to check the normality of the residuals/error terms. The residuals are plotted against a normal distribution in such a way that the points should form an approximate straight line. Departures from this straight line indicate departures from normality.

The importance of a Q-Q plot in linear regression is as follows:

1. Assumption Checking: One of the critical assumptions of linear regression is that the error terms are normally distributed. A Q-Q plot is used to visually check this assumption.
2. Outlier Detection: Q-Q plots are also a good way of identifying outliers in the data which might be skewing the regression analysis.
3. Model Comparison: Q-Q plots can also be used to compare two models and check which one gives a more normal distribution of residuals, hence is a better model.

   Remember, the Q-Q plot helps us to see at-a-glance how well our data fits the chosen distribution (in this case, normal distribution). It is not a formal statistical test and doesn't provide a p-value but gives a visual check about the normality assumption.