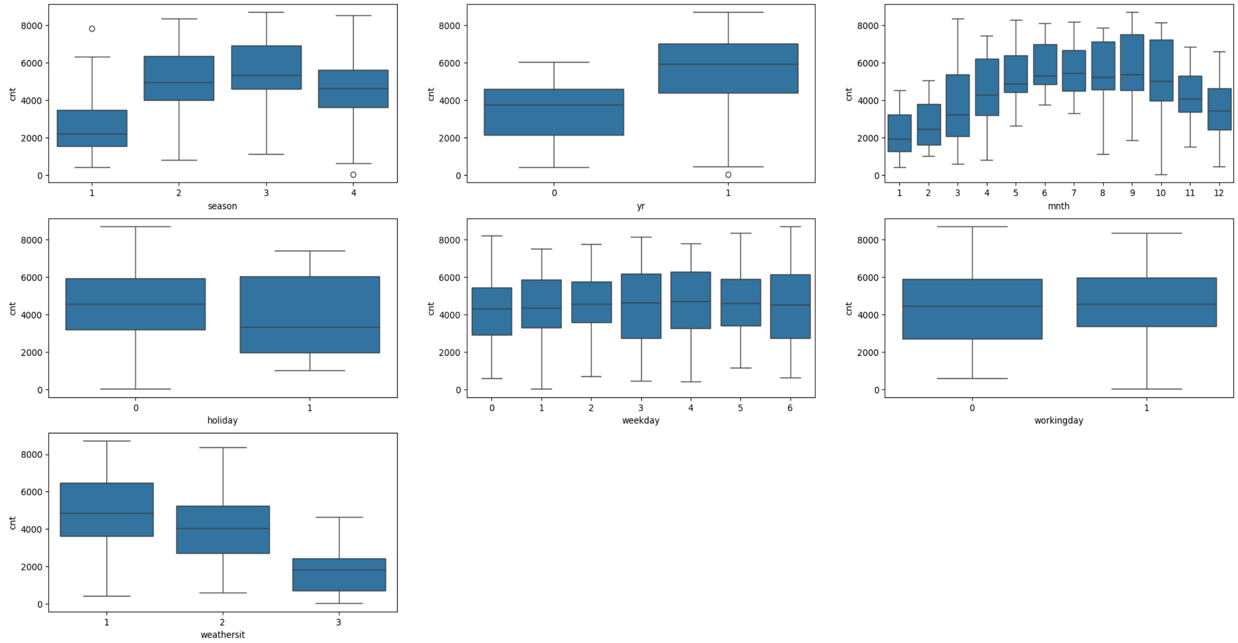
Assignment-based Subjective Questions

From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

I could identify the following variables as categorical variables in the dataset:
season, yr, mnth, holiday, weekday, workingday, weathersit
A boxplot was used to visualize these features. The following can be inferred from the plot:

- Season:
  - We observe an order fall, summer, winter, and then spring.
  - Median in fall approx 5000
- Year:
  - We can observe an increase in bike bookings in 2019 in comparison to 2018
  - Maybe because boombikes are getting more famous
- Month:
  - Bookings in months 4,5,6,7,8,9 and 10 observe that median is over 4000
  - So we observe a similar pattern in month and season
- Holiday:
  - Higher bookings when there is no holiday
- Weekday:
  - Almost the same median on all days of the week
- Working day:
  - Almost close to 5000 median bookings regardless of it being a holiday or not
- Weather situation:
  - Major booking when weather situation is Clear, followed by Mist + Cloudy then Light Snow
  - No booking when weather is Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog, which makes sense

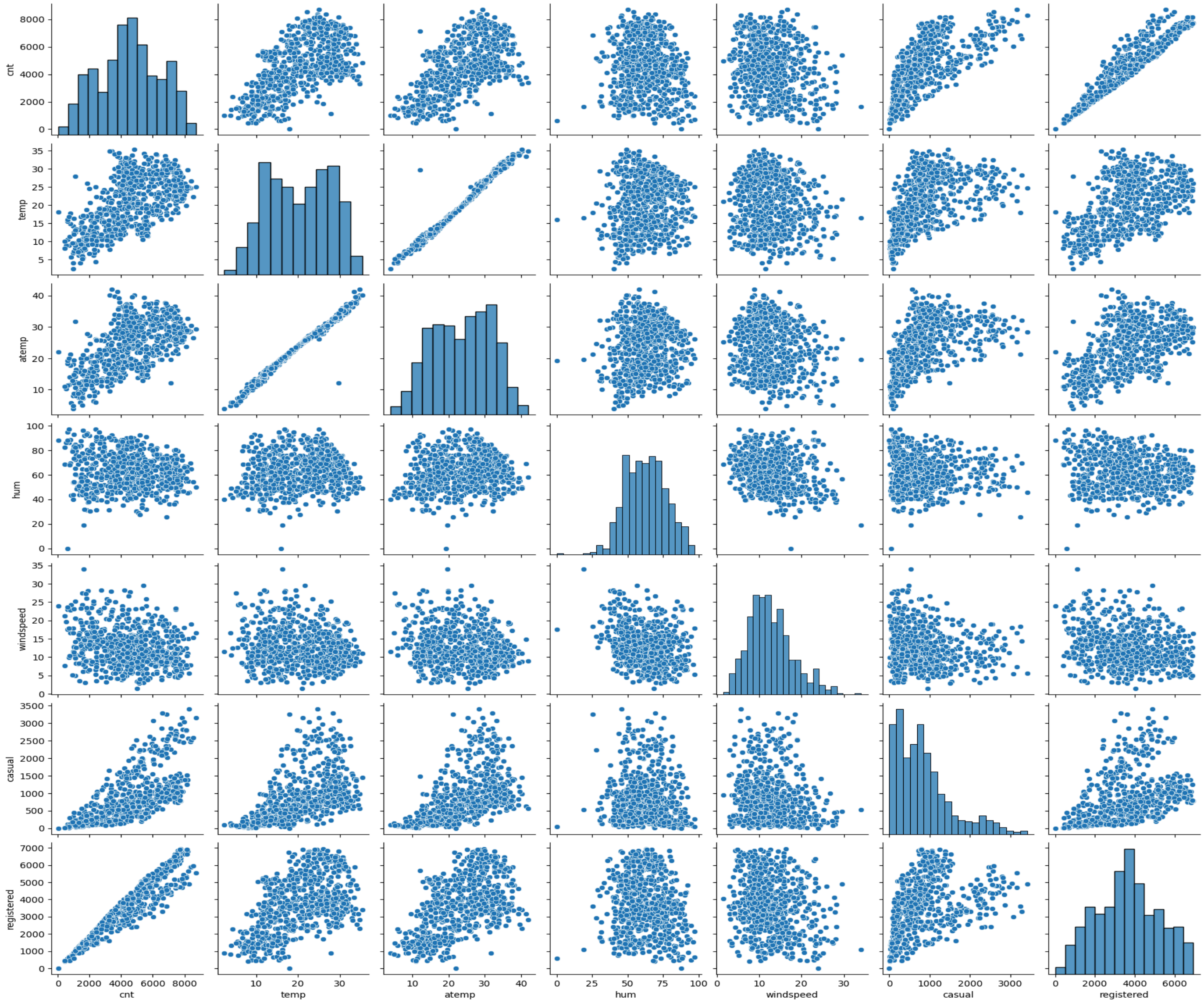


Why is it important to use drop_first=True during dummy variable creation?

- To get k-1 dummies out of k categorical levels by removing the first level
- We usually drop the first column because you can infer it from the other k-1 columns.
    - Eg:
        - Year had values 2018 and 2019,
        - and we converted 2019 and 2018 into 2 columns having boolean values 1 or 0.
        - And we deleted 2018 because we can get that information from 2019 column as follows if 2019 has a value 1 that means the bike was hired in 2019 or if 2019 column had a value of 0 that means the bike was hired in 2018

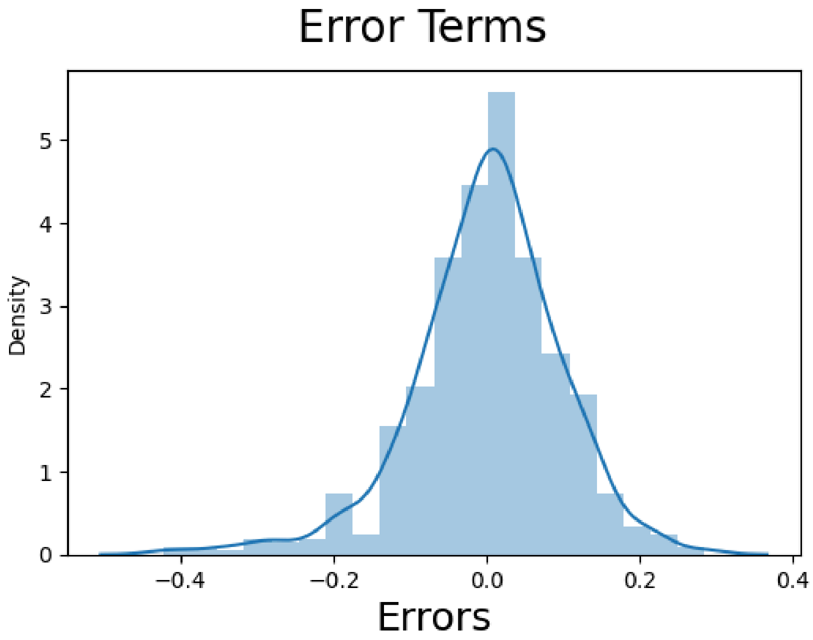
Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

If we don’t consider casual and registered, because they are part of ‘cnt’. Then ‘temp’ and ‘atemp’ have the highest correlation with the target variable ‘cnt’.



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

The distribution of error is normally distributed and centered around 0. The VIF amongst the predictors is below 5.



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

Based on the final model, the top 3 features contributing significantly are:

1. Temp:
   a. Co-efficient: 0.5158
   b. With a unit increase of temp, increases the cnt by 0.5158
2. Weather Situation 3:
   a. Co-efficient: -0.2872
   b. A unit increase in Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds, the bike hiring decreases by 0.2872
3. 2019
   a. Co-efficient: 0.2345
   b. We observed as year went by the increase in bike hire was by 0.2345

General Subjective Questions

Explain the linear regression algorithm in detail

Linear regression is a fundamental statistical method used to model the relationship between a dependent variable and one or more independent variables. It aims to find the best-fitting linear equation that describes how the dependent variable changes as the independent variables change. Here's a detailed explanation of the linear regression algorithm:

Simple Linear Regression

Simple linear regression involves one dependent variable y and one independent variable x. The goal is to find a linear relationship between x and y, which can be represented by the equation:

y = mx + c

where:

- y is the dependent variable, target variable, output
- x is the independent variable,
- c is the y-intercept of the regression line,
- m is the slope of the regression line,

Multiple Linear Regression

Multiple linear regression extends simple linear regression to include multiple independent variables. The equation becomes:

y = B0 + B1X1 + B2X2 + ….. + BnXn

where:

- y is the dependent variable,
- X1, X2… Xn are the independent variables,
- B0 is the y-intercept,
- B1, B2, …. Bn are the coefficients of the independent variables,

Assumptions of Linear Regression

1. **Linearity**: The relationship between the dependent and independent variables should be linear.
2. **Independence**: The observations should be independent of each other.
3. **Homoscedasticity**: The residuals (errors) should have constant variance at all levels of the independent variables.
4. **Normality**: The residuals should be approximately normally distributed.
5. **No Multicollinearity**: The independent variables should not be highly correlated with each other.

Explain the Anscombe’s quartet in detail.

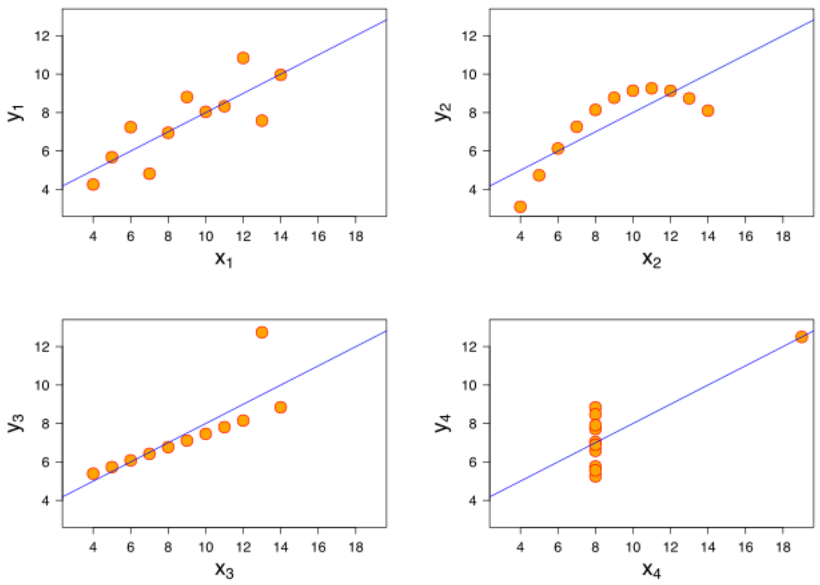
Anscombe's quartet is a set of four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. The quartet was created by the statistician Francis Anscombe in 1973 to illustrate the importance of graphing data before analyzing it and the effect of outliers and non-linearity on statistical measures.

Key Characteristics of Anscombe’s Quartet

Each of the four datasets in Anscombe’s quartet has the following properties:

- The same mean of x and y
- The same variance of x and y
- The same correlation between x and y
- The same linear regression line (y=3+0.5xy = 3 + 0.5xy=3+0.5x)
- The same coefficient of determination (R^2)

Despite these identical statistical properties, the datasets look very different when plotted. This emphasizes that relying solely on statistical properties without visualizing the data can be misleading.



Lessons from Anscombe’s Quartet

1. **Importance of Visualization**: Graphical representations can reveal patterns, trends, and outliers that are not apparent from summary statistics alone.
2. **Impact of Outliers**: Outliers can significantly affect the results of statistical analyses, such as regression coefficients and correlation.
3. **Non-Linearity**: Simple statistical measures like correlation and linear regression can be misleading in the presence of non-linear relationships.
4. **Data Integrity**: Always inspect the data visually to understand its structure and any peculiarities before relying on statistical summaries.

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 linear relationship between the variables. It ranges from -1 to 1. It is denoted by r

Between 0 and 1 indicates positive correlation
0 indicates no relation
Between 0 and -1 indicates negative correlation

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

Scaling is the process of adjusting the range and distribution of features in a dataset so that they fit within a specific range or follow a certain distribution. This is a crucial step in data preparation step as it can significantly impact the performance of the models. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

There are 2 types of scaling methods:

1. Normalized Scaling or Min-Max Scaling
    - Normalization (or min-max scaling) transforms the data to fit within a specific range, typically [0, 1]
    - Here outliers lie on 1 and 0
2. Standarized scaling
    - Standardization transforms the data to have a mean of 0 and a standard deviation of 1

When to Use Each Method:

- **Min-Max Scaling** is generally used when you need to constrain the values to a specific range. It is commonly used in image processing (pixel values) and neural networks where input values are typically expected to fall within a certain range.
- **Standardization** is preferred when the data has varying scales and you want to compare scores from different distributions. It is commonly used in linear regression, logistic regression, and many other machine learning algorithms.

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

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.
When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

I observed this when working on boombikes dataset, when I didn’t remove the variables ‘casual’ and ‘registered’ from the dataset. I noticed that the R2 of the model was 1 and the VIF was Infinity
After removing these 2 variables, did the model correct itself and dropped the R2 from 1 to 0.82

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

A Q-Q (quantile-quantile) plot is a graphical tool used to compare the distribution of a dataset with a theoretical distribution, such as the normal distribution. It helps assess whether the data follows a particular distribution by plotting the quantiles of the data against the quantiles of the theoretical distribution. If the data follows the theoretical distribution, the points on the Q-Q plot will approximately lie on a straight line. It allows you to inspect how much your distribution fits to a Gaussian, and where it deviates more.

Use and Importance of Q-Q Plots in Linear Regression

Q-Q plots are particularly useful in linear regression for the following reasons:

1. **Assessing Normality of Residuals**: Linear regression assumes that the residuals (errors) are normally distributed. A Q-Q plot of the residuals against a normal distribution can help check this assumption. If the residuals deviate significantly from the straight line, it suggests that the normality assumption is violated.
2. **Identifying Outliers**: Q-Q plots can help identify outliers in the data. Points that deviate significantly from the straight line may indicate outliers that can impact the regression model.
3. **Evaluating Model Fit**: By checking the distribution of residuals, Q-Q plots provide insight into whether the linear model is appropriate for the data. Deviations from normality in the residuals might suggest that a different model or a transformation of the data is needed.