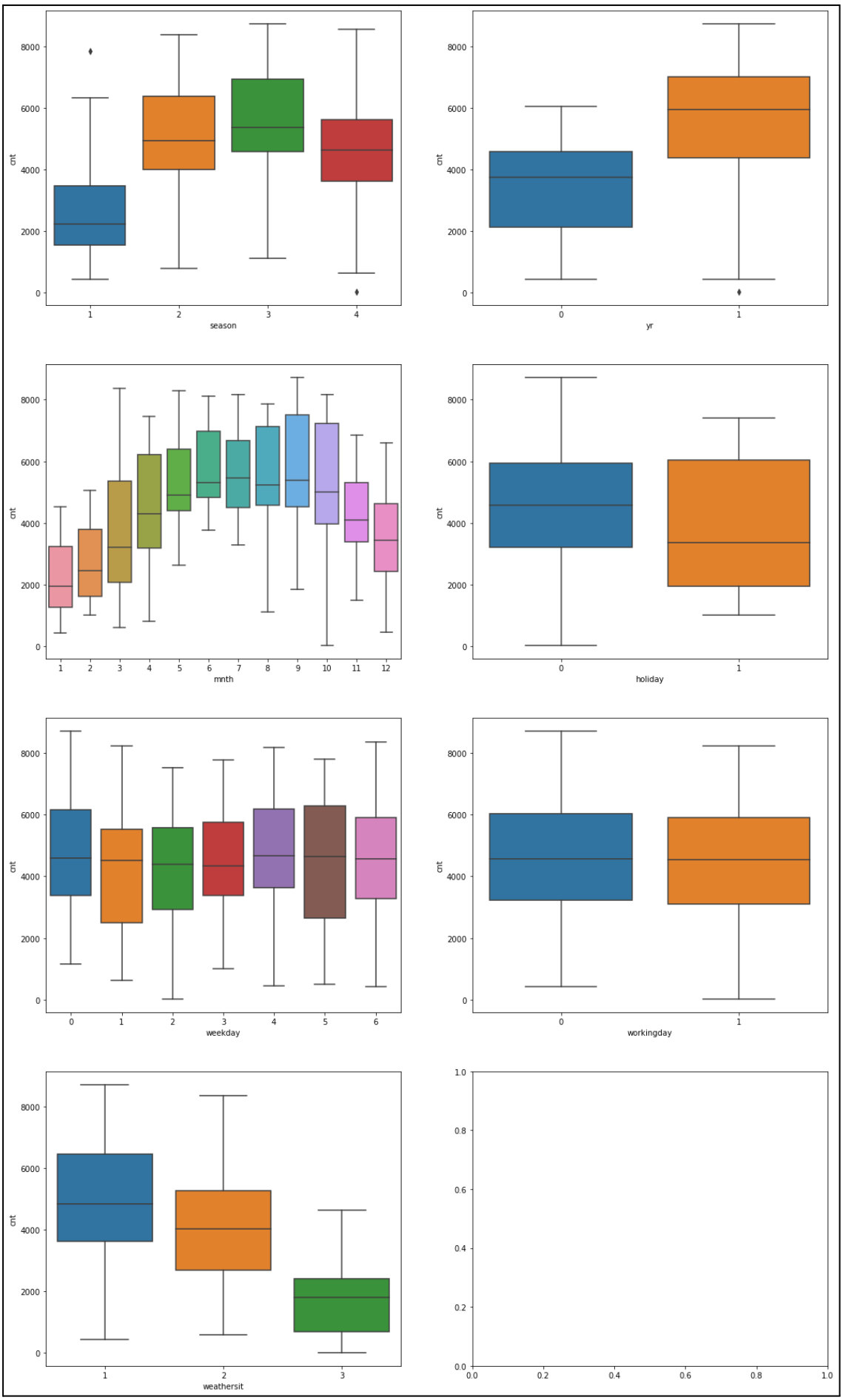
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
Ans. The relevant categorical variables in the dataset are 'season', 'yr' ,'mnth' ,'holiday' ,'weekday' ,'workingday', and 'weathersit' and different variables have different effects on the dependent variable.

- weekday has minimal impact, with similar demand on all days of the week
- workingday has minimal impact, with similar demand on working and non-working days
- season has moderate impact, with low demand during spring
- holiday has moderate impact, with lower demand on holidays
- yr has high impact, with significantly higher demand in 2019
- mnth has high impact, with higher demand between months 4 - 10
- weathersit has high impact, with lower demand for values 2 & 3

Box plots of the categorical variables are given on the next page.


cnt
season
yr
mnth
holiday
weekday
workingday
weathersit

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

Ans.

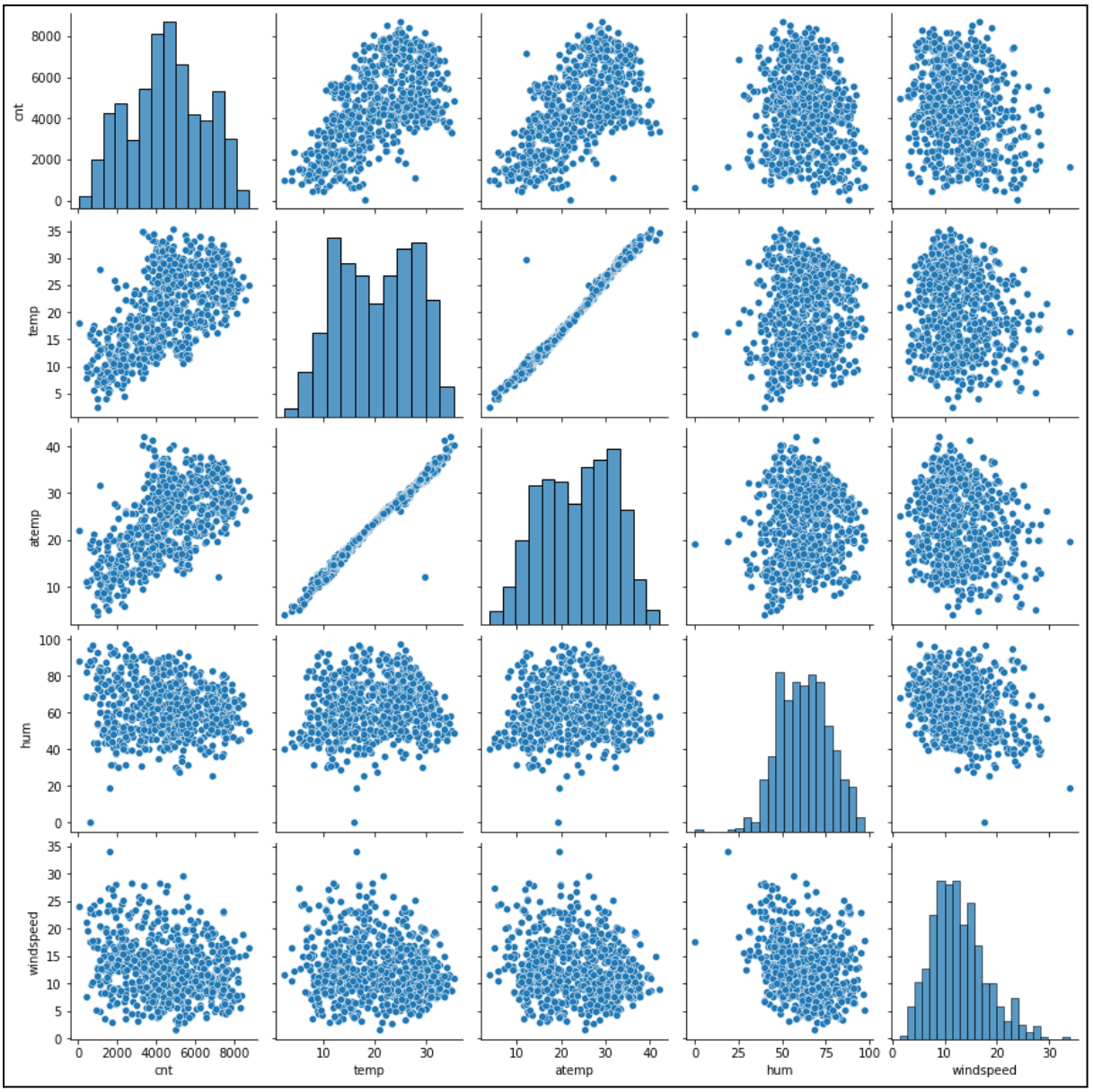
i) A categorical variable of p levels is encoded using p-1 dummy variables, but the Pandas get_dummies() function returns p dummy variables for such a categorical variable.

ii) Using drop_first=True with the get_dummies() function, drops one of the dummy variables and encodes the categorical variable using p-1 dummy variables.

iii) For example, using drop_first=True with the get_dummies() function for a categorical variable of 3 levels, returns 2 dummy variables.

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

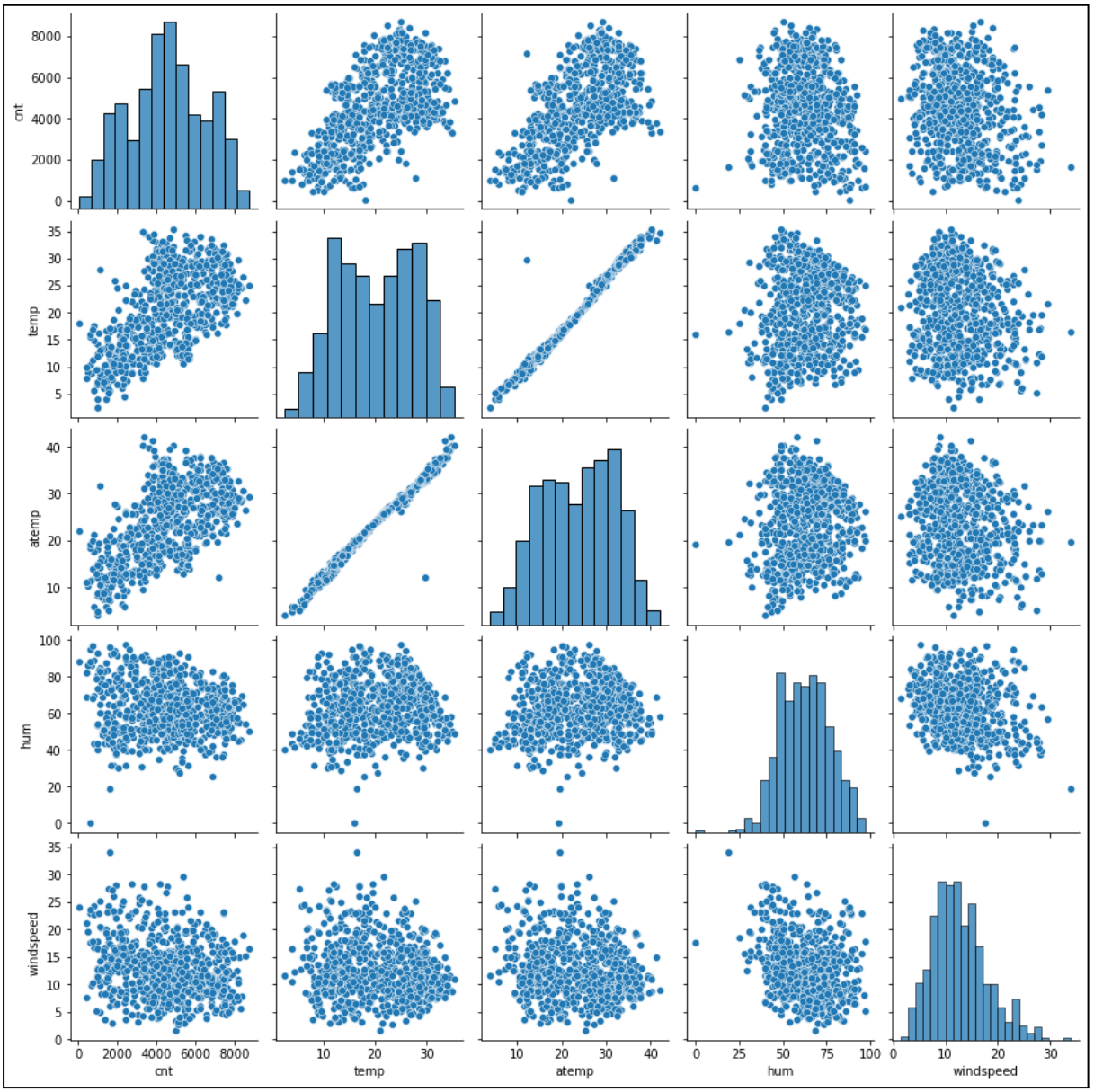
Ans. As seen from the pair plot, temp and atemp have the highest correlations with the target variable 'cnt'.



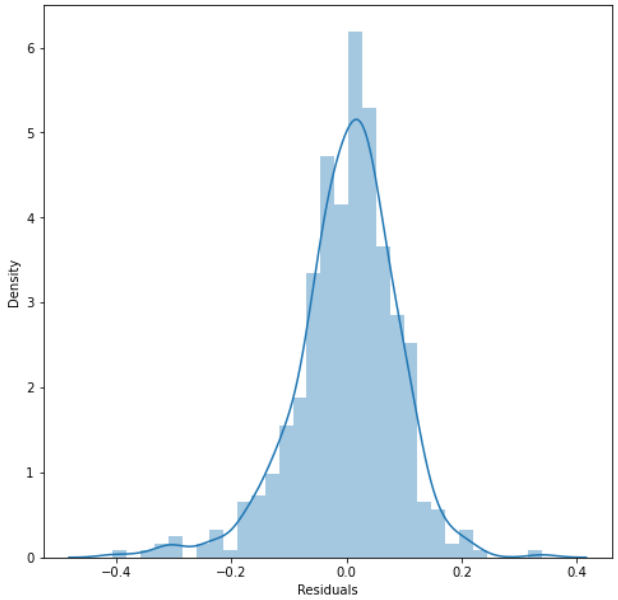
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans. I validated the assumptions of linear regression after building the model by plotting diagnostic charts as shown below.

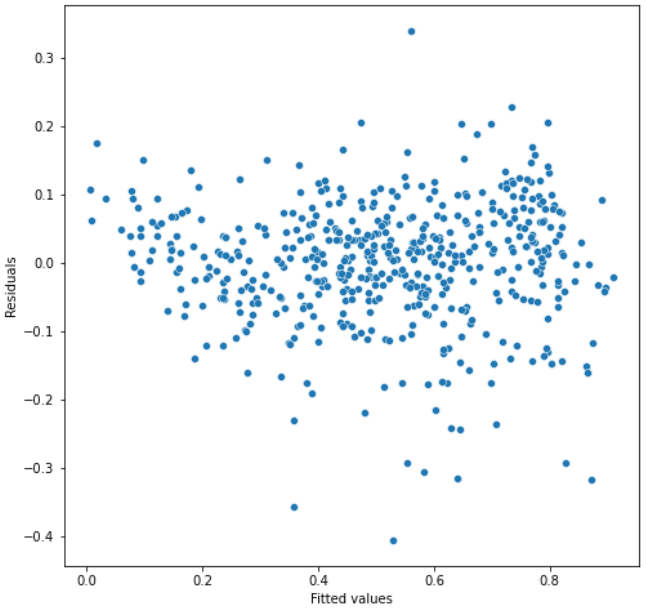
Linearity - Pair plot of target variable vs. numerical variables



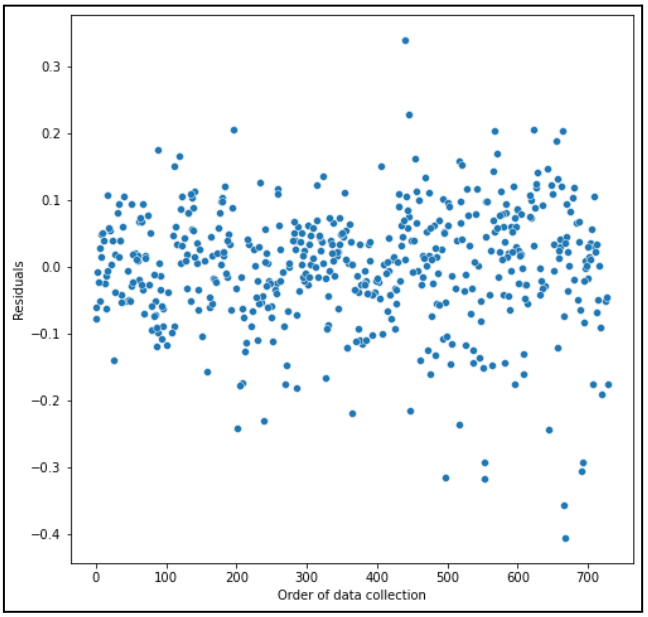
Normality - Histogram of residuals

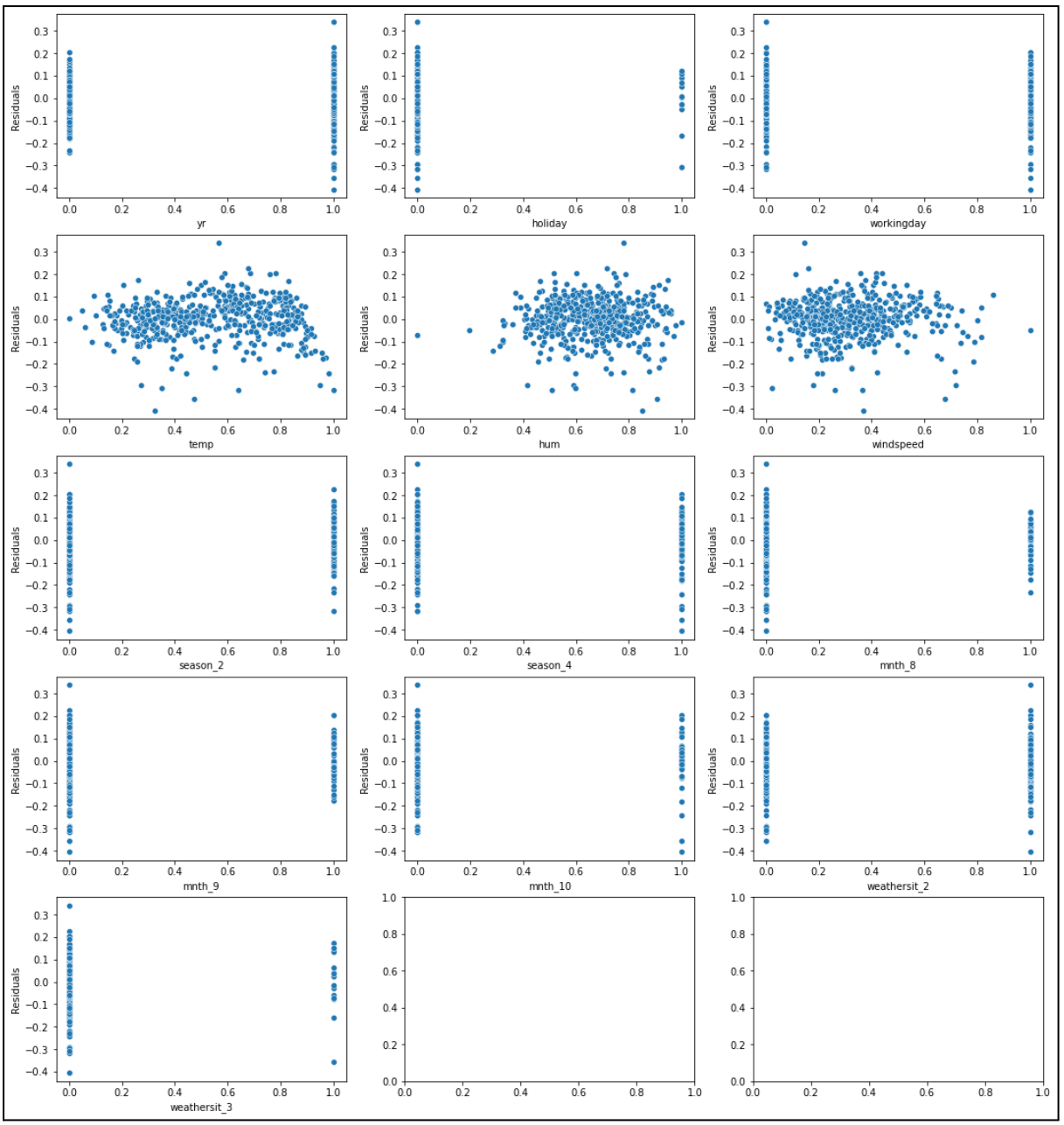


Homoscedasticity - Scatter plot of residuals versus fitted values



Independence - Scatter plots of (a) residuals vs. order of data collection and (b) residuals vs. each predictor




Residuals
yr
holiday
workingday
temp
hum
windspeed
season_2
season_4
mnth_8
mnth_9
mnth_10
weathersit_2
weathersit_3

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

Ans. Based on the final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are as follows:

- **temp** - when the normalised value of temp increases by 1, the normalised value of cnt increases by 0.5301
- **weathersit_3** - when weathersit_3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) takes the value 1 (compared to the reference 0), the normalised value of cnt decreases by 0.2407
- **yr** - when yr takes the value 1 (compared to the reference 0), the normalised value of cnt increases by 0.2289

General Subjective Questions

1. Explain the linear regression algorithm in detail.
Ans.

What Linear Regression Is
i) Linear regression is a machine learning algorithm that is used to express a response variable **y** as a function of linearly associated predictor variables $\mathbf{x_1,x_2,...,x_k}$.
ii) The linear function takes the form $\mathbf{y} = \beta_0 + \beta_1\mathbf{*x_1} + \beta_2\mathbf{*x_2} + ... + \beta_k\mathbf{*x_k}$ and the various β values are estimated using sample data.
iii) The case where **k = 1** is equivalent to fitting a line through the data.

Procedure
i) In order to develop the regression model, categorical variables are encoded as 1s and 0s using dummy variables, numerical data is often scaled, and the dataset is split into training and testing sets.
ii) The regression coefficients $\beta_0, \beta_1, \beta_2, ..., \beta_k$ are estimated by finding the values for which the sum of squares of difference between actual and predicted **y** values (called residuals) is minimum.
iii) This is done in software using the gradient descent algorithm, which iteratively tests β values to find the set corresponding to minimum residual sum of squares.
iv) Python libraries such as **statsmodels** and **scikit-learn** provide routines to estimate the β values and make predictions.
v) Backward elimination and forward selection are the 2 procedures used to select relevant predictor variables and arrive at an optimal model.

Assumptions of Linear Regression
i) A linear regression model can be built using a dataset when the following conditions hold.

- Linearity - There is a linear trend between the response and each of the predictor variables
- Normality - The residuals are normally distributed with mean 0
- Homoscedasticity - The variance of residuals is constant
- Independent observations - Successive observations must be independent

ii) The assumptions of linear regression are checked using appropriate diagnostic plots.

Model Fit & Prediction
i) The predictive power is measured using statistics such as R^2 and Adj. R^2, which indicate how well fitted values coincide with the actual values. Other measures such as F-statistic and AIC are also very useful.
ii) Once the model is finalized, it can be used for prediction, where **y** values are predicted for new sets of $\mathbf{x_1,x_2,...,x_k}$, or for inference, where $\beta_0, \beta_1, \beta_2, ..., \beta_k$ are used to identify the strength of the relationship between the predictor variables and the response variable.

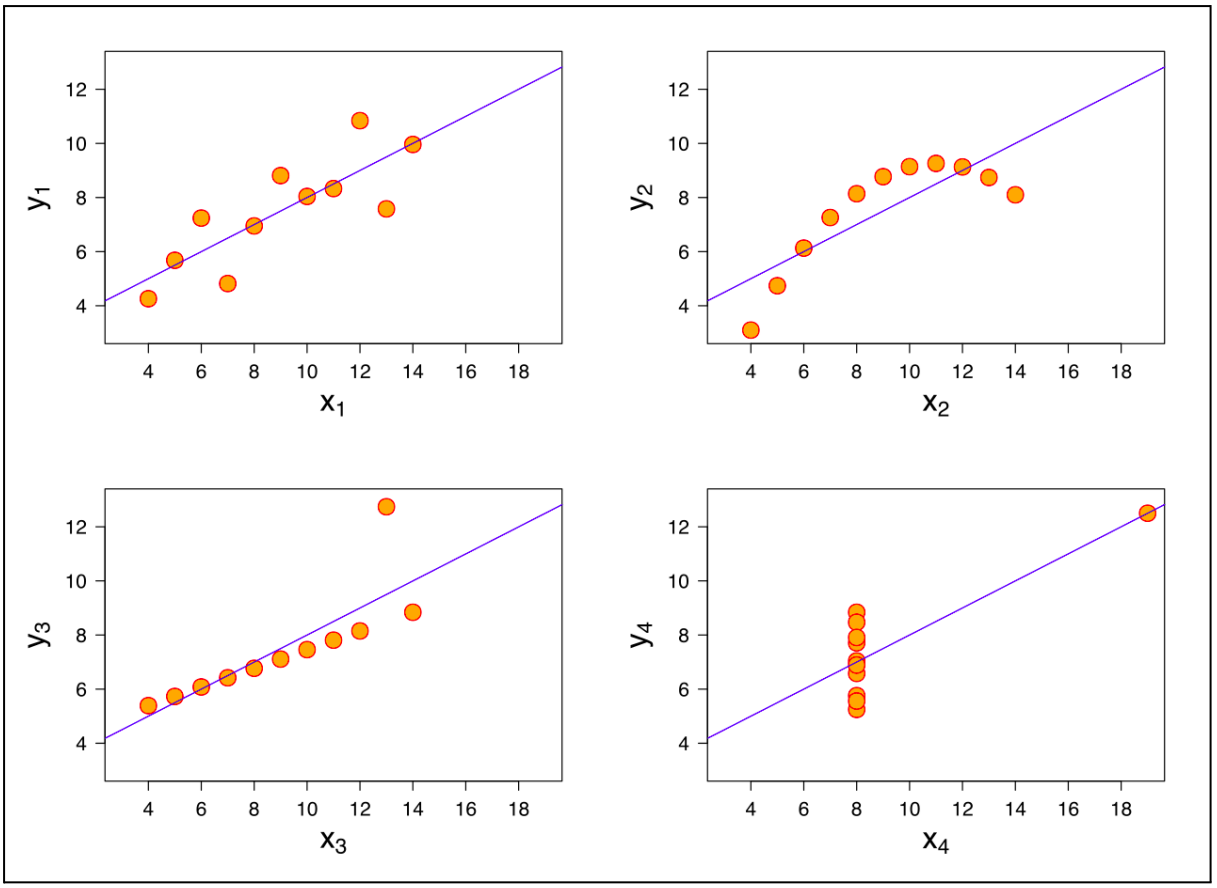
2. Explain the Anscombe's quartet in detail.

Ans.

i) Anscombe's quartet is a set of four data sets constructed by the statistician Francis Anscombe, which are nearly identical in many descriptive statistics but have different distributions and graphs.

ii) The nearly identical descriptive statistics are means of x & y, sample variances of x & y, correlation between x & y, linear regression lines and the coefficients of determination of linear regression.

iii) Scatter plots of the datasets are shown below.



iv) Dataset 1 features a simple linear relationship between x and y

Dataset 2 features a nonlinear relationship between x and y

Dataset 3 features a simple linear relationship between x and y with one outlier

Dataset 4 features a simple linear relationship between x and y with one high-leverage outlier

3. What is Pearson's R?

Ans.

i) Pearson's R or the Pearson correlation coefficient of 2 variables is equal to their covariance divided by the product of their standard deviations.

ii) Pearson's R is a measure of linear correlation between the 2 variables and takes values between -1 and 1.

iii) A positive R close to 1 indicates a positive linear relationship between the variables and a negative R close to 1 indicates a negative linear relationship.

iv) Pearson's R is only a measure of the linear relationship between variables and does not factor in nonlinear relationships.

v) For example, the variables 'height' and 'weight' of a sample of adults would have a high positive correlation.

vi) The formula for Pearson's R is as follows:

$$\rho_{X,Y} = \frac{\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

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

Ans.

i) Scaling is a data preprocessing step in machine learning, wherein values of variables are mapped to another scale.

ii) Scaling is performed for reasons such as reducing the effect of outliers, better interpretability, faster convergence of algorithms, etc.

iii) Normalization and standarization are 2 types of scaling and their differences are given below.

Normalization	**Standardization**
Normalization is used to map variables to a similar scale.	Standardization transforms variables by subtracting from mean and dividing by standard deviation.
The scaled value is given by X_new = (X - X_min)/(X_max - X_min)	The scaled value is given by X_new = (X - mean)/Std
Range of scaled values is [0,1] or [-1,1].	Not limited to a specific range.
It is used when features are of different scales.	It is used to ensure zero mean and unit standard deviation.
It is affected by outliers.	It is not affected by outliers.

MinMaxScaler from Scikit-Learn is used for transformation.	StandardScaler from Scikit-Learn is used for transformation.
Useful when the distribution is unknown.	Useful when the distribution is normal.

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

Ans.

i) Consider a multiple linear regression model with k predictor variables $\mathbf{x_1}$,$\mathbf{x_2}$,...,$\mathbf{x_k}$ and a response variable **y**.

ii) The Variance Inflation Factor $\mathbf{VIF_i}$ of a predictor variable $\mathbf{x_i}$ is given by the following formula, where $\mathbf{R_i}$ is the coefficient of determination of the linear regression model with variable $\mathbf{x_i}$ as the target variable and the remaining predictor variables as predictors.

$$VIF_i = \frac{1}{1 - R_i^2}$$

ii) $\mathbf{VIF_i}$ can take the value infinity when $\mathbf{R_i}$ is 1, which happens when $\mathbf{x_i}$ is perfectly fitted by a linear regression model with the remaining predictor variables i.e. residual sum of squares is zero.

iii) A VIF value of infinity indicates severe multicollinearity and the associated variable is usually removed from the original regression model.

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

Ans.

i) A Q-Q plot or quantile-quantile plot is a scatter plot of specified quantiles of one variable against corresponding quantiles of another variable.

ii) For example, a scatter plot of each percentile value of a variable **x** against each percentile value of another variable **y** would be a Q-Q plot of **x** and **y**.

iii) The Q-Q plot is used to check whether the 2 variables have the same distribution, in which case the plot would be a straight line. Thus, the Q-Q plot can be used to identify the underlying distribution of a certain variable, such as exponential, normal, etc.

iv) In linear regression, the Q-Q plot is used to determine whether the training and testing datasets are from the same population.

v) The main advantages of using a Q-Q plot in linear regression are that the following can be detected:

- Whether both datasets have similar distributions.
- distributional aspects like shift in location, shift in scale, changes in symmetry and the presence of outliers,
- location and scale of variables, and

- whether both variables have the same tail behaviour.

vi) An example of a Q-Q plot in which the data is normally distributed is given below.