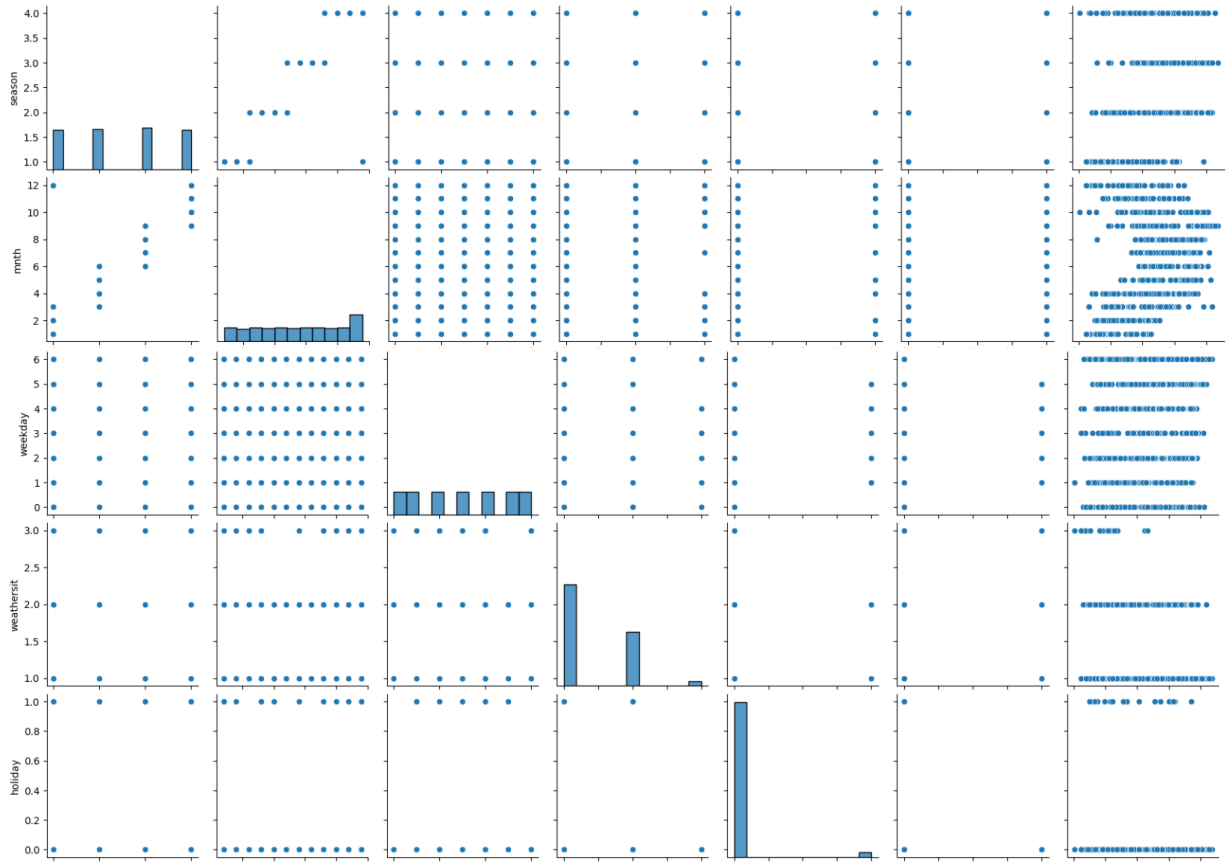
Assignment-based Subjective Questions

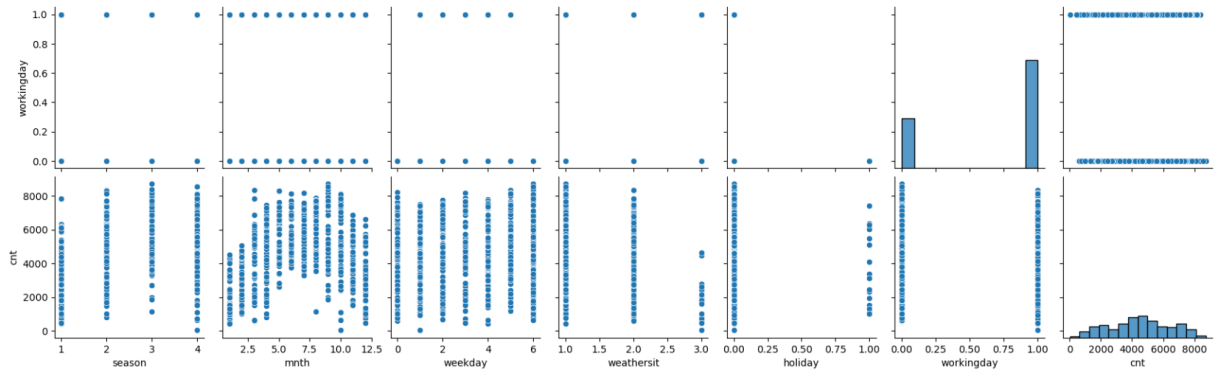
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer:

I have considered 'CNT' as dependent variable and the following as categorical variable:

1. **Season** [Inference after analysis: Summer and winter has positive correlation with count- indicating Boombike can keep more bikes in the streets. Spring has a negative Correlation with count indicating lesser requirements]
2. **Month** [Inference after analysis: September has a positive correlation with count while Januray and July are negatively correlated]
3. **Weekday**[Inference after analysis: non-significant correlation with count]
4. **WeatherSit**[Inference after analysis: value 3(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) and 4 (Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog) has negative correlation with count]
5. **Holiday**[Inference for analysis: Holdiay has a negative correlation with count. People are not preferring holidays for cycling for some reason. Working days are more preferred.]
6. **WorkingDay**[Inference for analysis: non-significant correlation with count and explained by holiday]





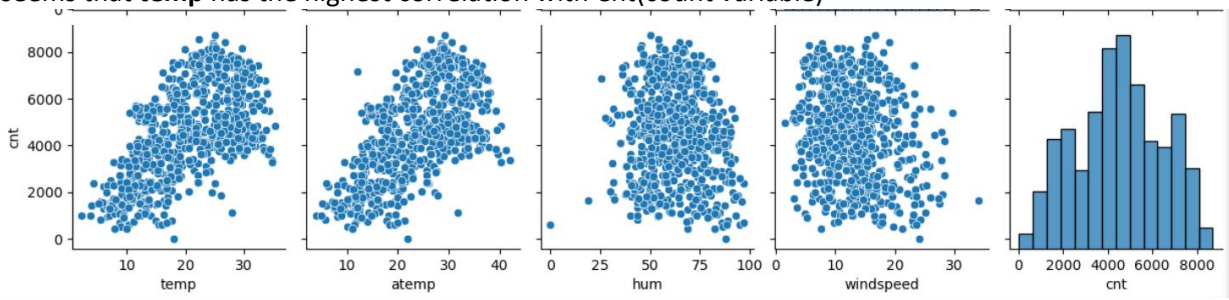
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Answer: Because the first column can be deduced from the combined values of remaining columns/variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer:

It seems that **temp** has the highest correlation with Cnt(count variable)



4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer:

Using the R-squared score on the test set we derived a value : 0.8025684920603356. This validates our assumptions.

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:

From the linear Regression model we got the following absolute coefficients

temp 0.472115
Snowy 0.285425
yr 0.234283
windspeed 0.154916

From above we can say that "**temp**", "**Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds**" and "**year**" contributes significantly towards demand

General Subjective Questions

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

Answer:

Key Legend:

1. **Definition**: A linear regression model with multiple predictors is expressed by:

$y=\beta 0+\beta_1 x_1+ \beta_2 x_2 +\beta_3 x_3 +\ldots+ \beta_n x_n +\epsilon$
where x_n are predictor variables and β_n are the corresponding coefficients.

2. **Model fitting**: The goal is to estimate the coefficients so as to minimize the residual sum of squares (RSS)
3. **Method:** Ordinary Least Squares is the most common method for fitting a linear regression model. It minimizes the sum of the squared residuals.
4. **Evaluation:** Checking adjusted R square for test and train data

With the above legend we can start explaining the process of target variable prediction using independent predictors

1. **Data understanding and data loading :** The data provided should be loaded as a python data frame and data dictionary should be referred to understand the business problem represented by that data
2. **Pre-processing steps:**
    a. Drop any variable(column) which is evidently non-contributing like serial numbers/date in current case
    b. Map all categorical variables from data dictionary
    c. Create dummy variables for all categorical variables where no. of values are more than 2 and drop the first so as to have no. of values -1 dummy variables(columns)
3. **EDA: Perform data analysis**
    a. Univariate Analysis
    b. Bivariate Analysis
4. **Train-test split**: Split the total dataset into training set and test set. The idea is to develop and fine tune the model using training data set. Predict the training data using the model and select the variables based on their p values, VIF values and conclude on a good r square value reported by the model. Later use the model to predict target test data set.
    - Usually we are doing a 70:30 split with random set 100
5. **Missing value imputation:** Impute the missing values using the **mode** of the column values. If most of the values in that column are missing, drop the same.
6. **Scaling:** scale the numerical data so as to express them in same range. We can use the following methods:
    a. **Standardization**
    b. **Min-max(normalized):** We are using the min-max to express all of the data concentrated to the range of 1 to 0. This is not necessary for the categorical variable as they are already being represented by 0/1 value in dummy variables or in original format.

    Scaling can be implemented using two functions of scalar object:
    a. **Fit_transform:** this is used on training data set.
    b. **Transform:** This is used on test data set.
7. **Feature Selection:** Next is feature selection. We decide on which variables best explains the r-square. Starting from considering all variables, we remove variables one-by-one**.** This can be done in couple of ways:
    a. **RFE:** Automatic selection of columns. We can set the number of features we would like to use to evaluate the r-square
    b. **Manual:** We do the above work manually and repeat till we are satisfied with the r square value. This is done by elimination of high p-value variables and >=5 VIF values.

Sometimes we might keep back the high VIF value variables if business knowledge wise it makes sense to keep those back. But usually not for very high VIF value variables.

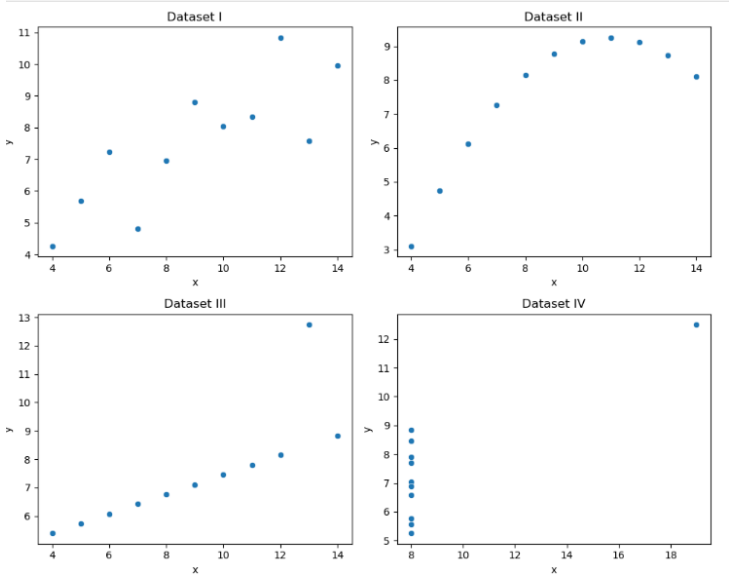
c. **Hybrid:** In case of a lot of columns, we also use both the above option in tandem to reduce effort and also use manual understanding of the business knowledge. We first let RFE reduce the number of features. Then use maual picking to keep the more menaing full ones and remove less significant ones. This part still used p-value/VIF valuation process for each variable and removing on-by-one accordingly.

8. **Model Building:**
    a. Separate out the target variable from rest of them and create the y_train for target variable and x_train for rest
    b. Using statsmodels.api, sklearn.LinearModel we create the model and fit the model to the x train and y train
    c. Readjust with new set of columns from output of the above using model's summary method which lists down the R-square, p-Value.
    d. Repeat the model fit with new columns after eliminating columns that has highe p-Value(>.05)
    e. After satisfaction start with the VIF calculation of remaining variables using statsmodel.
    f. **VIF**: Variance inflation factor: This means how much a variable can be explained by other variable values. The more the VIF the less important is the parameter. Hence we remove this too. The usual threshold is to remove anything more than 5. variance_inflation_factor is used for the same its provided in statsmodel module.
9. **Evaluation**:
    a. This is final step where to use the model to predict the test data.
    b. Calculate the r square with the test data's predicted y value i.e, target variable and actual target variable using the variables in our model.

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

Answer:

By examining Anscombe's quartet, we learn that relying solely on summary statistics can lead to incorrect conclusions, and visualization is an indispensable tool in data analysis.



Anscombe's quartet is a set of four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. It was created by the statistician Francis Anscombe to demonstrate

the importance of graphing data before analyzing it and to show how different datasets can have the same statistical properties but different distributions and relationships.

- Each of the four datasets in the quartet consists of 11 (x, y) points.

All four datasets have the same but the visual distribution is very different as shown in above diagram:

a. Mean of x
b. Mean of y
c. Variance of x
d. Variance of y
e. Correlation coefficient between x and y
f. Linear regression line: y=3.00+0.500x
g. Coefficient of determination (R2) of the linear regression

Anscombe's quartet underscores the importance of:

Visualizing Data: Before performing any statistical analysis, visualizing data uncovers patterns, outliers, and relationships that summary statistics might not reveal.
Exploratory Data Analysis (EDA): A critical step in data analysis that involves visualizing and exploring data to inform subsequent statistical modeling and hypothesis testing.

3. What is Pearson's R? (3 marks)
Answer:
Pearson's R, also known as the Pearson correlation coefficient or simply the correlation coefficient, is a measure of the linear relationship between two variables. It quantifies the degree to which a change in one variable predicts a change in another variable.

Positive Correlation ($0<R\leq1$): As one variable increases, the other variable tends to increase.
Negative Correlation ($-1\leq R<0$): As one variable increases, the other variable tends to decrease.
No Correlation ($R=0$): No linear relationship between the variables.

In our assignment , **Summer and winter** has **positive** correlation with number of bike shares and **Spring** has a **negative** Correlation with count

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
Answer:
Scaling: We scale the numerical data so as to express them in same range.In machine learning this helps the models to perform better and converge to optimal parameters with better accuracy. We can use the following methods:

a. **Standardization**: In machine learning, this is a pre-processing technique where the data variables are transformed so as to have a mean of zero and a standard deviation of one. In this process we find the average (mean) of your data. Subtract this average from each data point. Divide the result by the standard deviation (a measure of how spread out the data is). One effect of Standardization is that it impacts even categorical variables.
b. **Min-max(normalized):** We are using the min-max to express all of the data concentrated to the range of 1 to 0. This is not necessary for the categorical variable as they are already being represented by 0/1 value in dummy variables or in original format.

Min max Scaling can be implemented using two functions of scalar object:

a. **Fit_transform**: this is used on training data set.

b. **Transform:** This is used on test data set.

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

Answer:

VIF is expressed as $1/(1-R_i^2)$

This means the more the R square approaches 1 the denominator becomes very small and dividing 1 with smaller denominators makes bigger value for VIF tending to infinity as R square grows. In modeling this is means the more the Rsquare the more a variable can be expressed by other variables thus making the variable itself unnecessary to used in model.

An R square =1 means the target can be completely expressed by another variables. In our assignment the registered and casual variables are such which can express target variable completely. So we have not considered them.

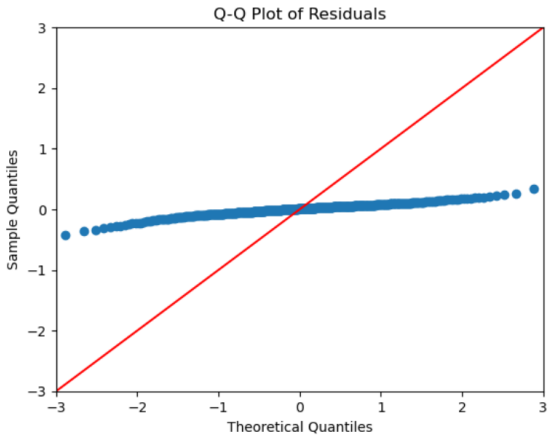
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

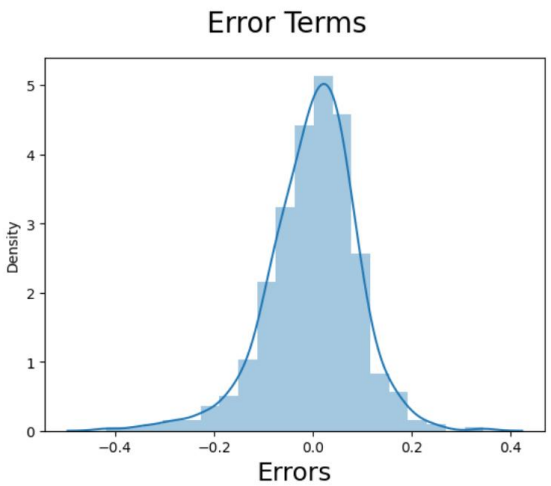
Answer:

A Q-Q (quantile-quantile) plot is a graphical tool used to assess whether a set of data follows a particular theoretical distribution, typically the normal distribution. It plots the quantiles of the sample data against the quantiles of the theoretical distribution. If the data comes from the specified distribution, the points on the Q-Q plot will approximately lie on a straight line.

In our Bike sharing assignment we QQ plotted the residuals:

```
[53]: sm.qqplot(residuals, line='45')
      plt.title('Q-Q Plot of Residuals')
      plt.show()
```





We can observe that the Sample Quantile follows a roughly straightline with same inclination direction. The error has normal disctribution.

Some important concepts:

Quantiles: Quantiles are cut points that divide the range of a probability distribution into continuous intervals with equal probabilities. For example, the median is the 0.5 quantile.

Plotting: The quantiles of the sample data are plotted against the quantiles of the theoretical distribution. If the sample data is from the theoretical distribution, the points should fall along a reference line (usually the 45-degree line).
Interpreting a Q-Q Plot:

Linear Relationship: If the points form a roughly straight line, the data follows the theoretical distribution.
Deviations from Linearity: Systematic deviations from the straight line indicate departures from the theoretical distribution. For instance:
Heavy Tails: Points deviate upwards at the ends of the plot.
Light Tails: Points deviate downwards at the ends of the plot.
Skewness: Points form an S-shape.

It helps :
To visually assess whether the data follows a specific theoretical distribution.
To identify deviations from normality, such as skewness and kurtosis.
To detect outliers.

How It Works:

Use and Importance in Linear Regression
In linear regression, one of the key assumptions is that the residuals (errors) are normally distributed. A Q-Q plot can be used to assess this assumption:

Checking Normality of Residuals:

After fitting a linear regression model, a Q-Q plot of the residuals can be created.
If the residuals follow a normal distribution, the points on the Q-Q plot should lie approximately on the 45-degree reference line.
Identifying Problems:

Non-Normality: If the residuals deviate significantly from the line, it indicates that the residuals are not normally distributed.
Heteroscedasticity: Patterns or systematic deviations can indicate heteroscedasticity (non-constant variance of residuals).
Outliers: Points far away from the reference line indicate potential outliers.
Importance:

Model Validity: Ensuring that residuals are normally distributed validates the model assumptions and supports the reliability of statistical tests (e.g., confidence intervals and hypothesis tests).
Model Improvement: Identifying deviations from normality can suggest the need for data transformation or alternative modeling approaches.