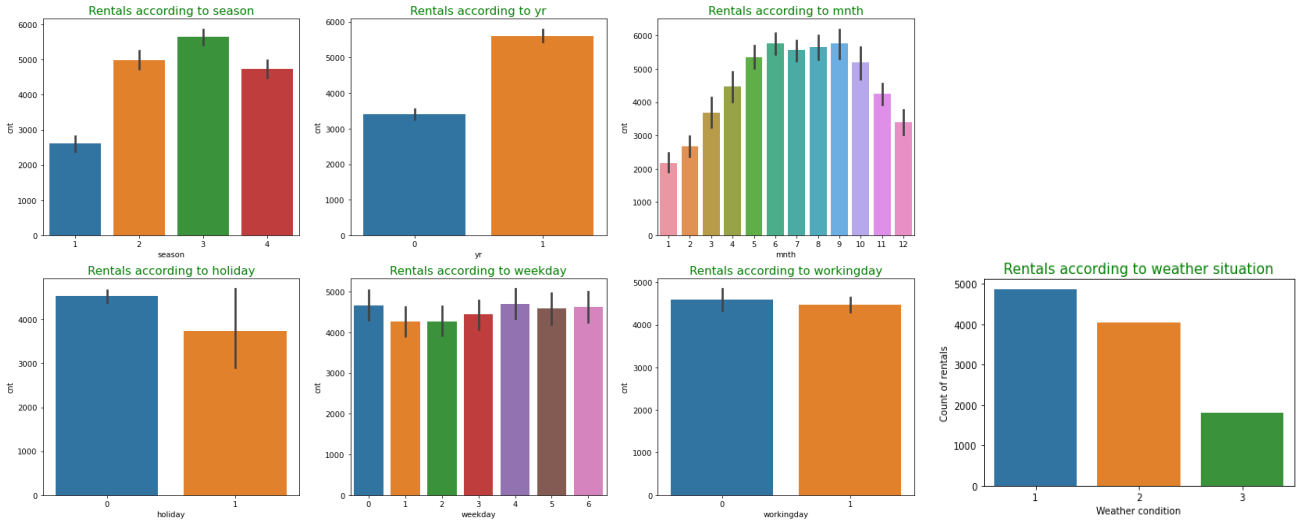
ASSIGNMENT

Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

1. Weather:
    - We can see that the greatest number of rentals are on days with clear skies, partly cloudy or misty skies
    - Very few rentals on days with light rains/snowy days
    - No rentals on days with Heavy rains
2. Season:
    - Highest average rentals in Fall season and lowest in Spring
3. Year:
    - Average Rentals are more in 2019 than 2018
4. Month:
    - Average Rentals are high in June-Sept, lowest in Jan-Feb
5. Holidays:
    - Average Rentals are high on Non-holidays
6. Weekdays:
    - Average rentals are high during weekends and Thursday, Friday
7. Working days:
    - Demand is similar irrespective of working days or not



Q2. Why is it important to use drop_first=True during dummy variable creation?

During dummy variable creation, for a variable with ‘p’ levels, ‘p’ new dummies are created. However, one of these dummy variables can be inferred using the other p-1 dummy variables. So, we should drop one of them and drop_first=True does that for us.

For Eg: If we have a column as Weather, with 3 levels: Hot, Cold, Rainy.

Date	Temperature	Weather
01-12-2020	5	Cold
01-02-2021	12	Cold
01-04-2021	32	Hot
01-07-2021	26	Rainy

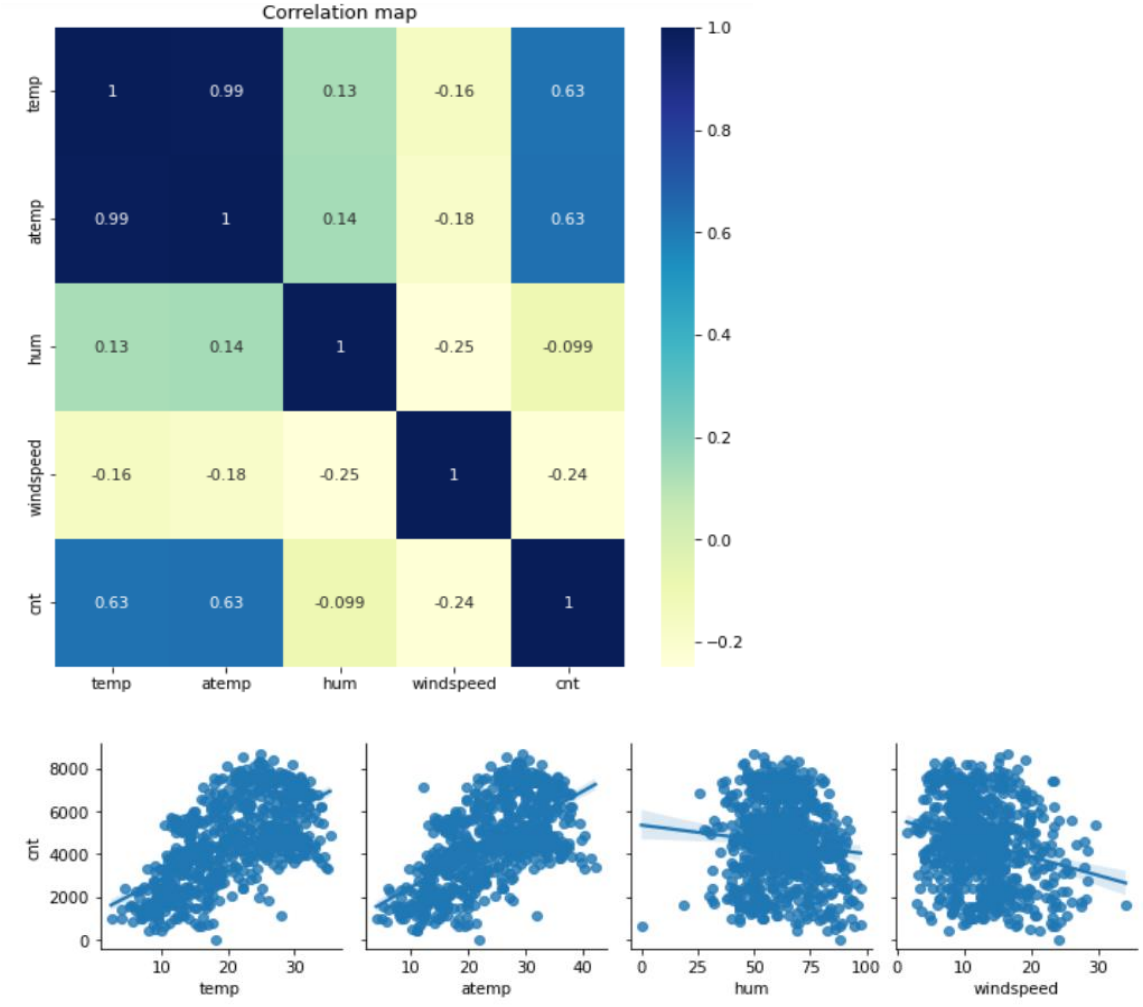
When we create dummies for Weather:

Date	Temperature	Weather	Hot	Cold	Rainy
01-12-2020	5	Cold	0	1	0
01-02-2021	12	Cold	0	1	0
01-04-2021	32	Hot	1	0	0
01-07-2021	26	Rainy	0	0	1

Here, even if we remove Hot variable, Cold and Rainy are enough to infer the Weather as only one of them can be 1 for a row.

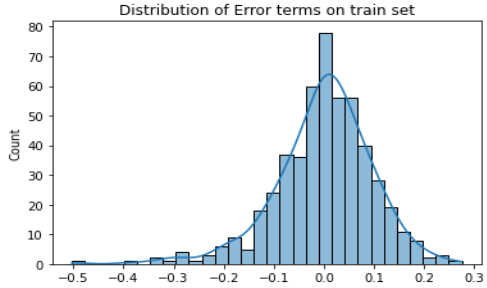
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Temperature has the highest correlation with the target variable (cnt).

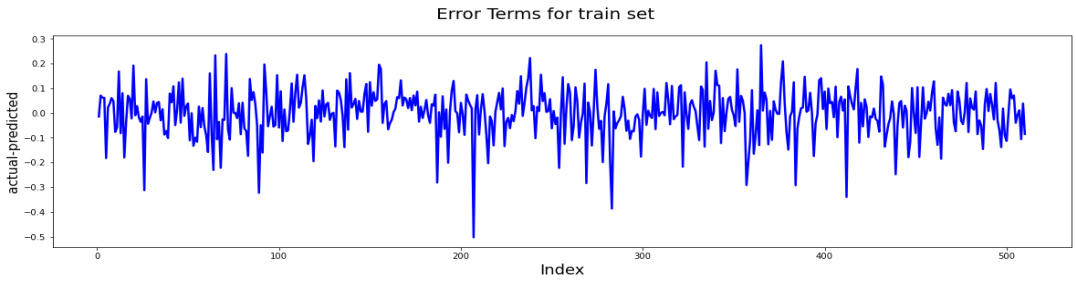


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

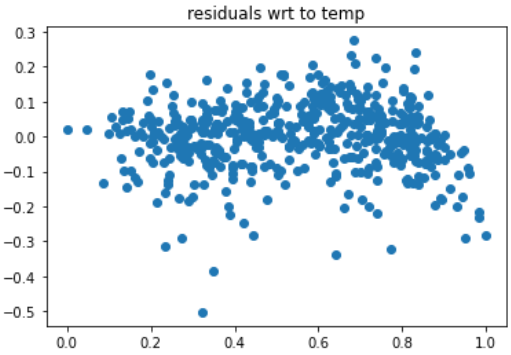
- The residuals are normally distributed with mean zero.

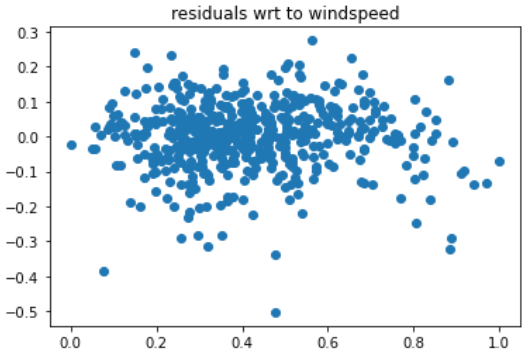


- The error terms are random in nature and no patterns are present. This means that the points are independent of each other.



- Homoscedasticity check: The residuals have a constant variance wrt independent variables.





- No multicollinearity between variables as VIF scores are <= 5.

	Features	VIF
1	temp	5.00
2	windspeed	5.00
6	winter	2.35
0	yr	2.08
5	spring	1.80
10	Nov	1.76
4	Misty	1.57
8	Jul	1.38
7	Dec	1.32
9	Mar	1.16
3	Light_rain	1.10

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

The top 3 features are:

1. Temp
2. Light_rain
3. yr

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Linear regression is a supervised learning method used to model the relationship between a dependent variable and one or more independent variables. In simple linear regression, there is only one independent variable, while in multiple linear regression, there are multiple independent variables.

The equation of best fit line for a dependent variable y wrt independent varibles $X_0, X_1, \ldots, X_n$ is:

$y_{pred} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots. + \beta_n X_n$

Where, β_0 = intercept and $\beta_1, \ldots., \beta_n$ are coefficients of $X_1, \ldots, X_n$ respectively.

The model accuracy is evaluated using the Residuals which are calculated as follows:

Residual = $y - y_{pred}$

These residuals might be positive or negative, so we take their squares to get an idea of the total errors in prediction:

Residual Sum of Squares (RSS) = $\sum^{n}_{i=1} (y_i - y_{pred})^2$

Assumptions of Linear Regression:

There are several assumptions that must be met for linear regression to be valid. Violation of these assumptions can lead to biased and unreliable results. The assumptions are:

1. Linearity: The relationship between the dependent variable and the independent variables must be linear.
2. Independence: The observations must be independent of each other. This means that the value of one observation should not be influenced by the value of another observation.
3. Homoscedasticity: The variance of the errors should be constant across all values of the independent variables. In other words, the spread of the residuals should be the same for all values of x.
4. Normality: The errors should be normally distributed. This means that the residuals should follow a normal distribution.
5. No Multicollinearity: The independent variables should not be highly correlated with each other. High levels of multicollinearity can lead to unstable estimates of the regression coefficients.

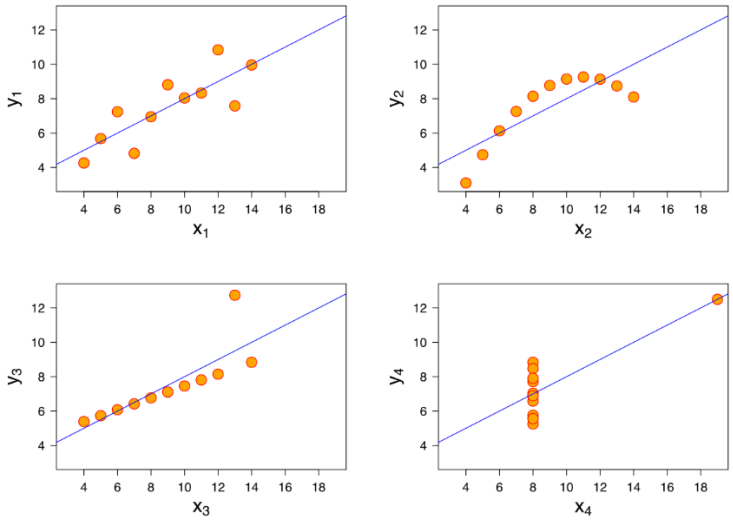
Q2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a set of four datasets, each containing 11 (x,y) data points. The datasets are constructed in such a way that they have nearly identical statistical properties, yet they look very different when graphed. The quartet was created by the statistician Francis Anscombe in 1973 to illustrate the importance of visualizing data, rather than relying solely on summary statistics.

The four datasets in Anscombe's quartet have the following properties:

1. In the first dataset, the relationship between x and y is linear. The correlation coefficient between x and y is 0.816, and the regression line y = 0.5x + 3 is a good fit for the data.
2. In the second dataset, the relationship between x and y is nonlinear. The correlation coefficient between x and y is 0.816, but the regression line y = 0.5x + 3 is an even worse fit.
3. In the third dataset, the relationship between x and y is also linear, but with a different slope. The correlation coefficient between x and y is again 0.816, but the regression line y = 0.5x + 3 is a poor fit for the data because there is one outlier that pulls the line away from the main cluster of points.
4. In the fourth dataset, the relationship between x and y is again linear, but with a strong outlier that has a large influence on the regression line. The correlation coefficient between x and y is 0.816, but the regression line y = 0.5x + 3 is again a poor fit for the data.

By creating these four datasets, Anscombe demonstrated that it is important to visualize data in order to detect patterns, outliers, and other features that summary statistics may miss.



Q3. What is Pearson's R?

Pearson's correlation coefficient, also known as Pearson's R, is a measure of the linear correlation between two variables. It is a statistical technique used to determine the strength and direction of the linear relationship between two variables.

The value of Pearson's R ranges from -1 to +1, with -1 indicating a perfect negative correlation, 0 indicating no correlation, and +1 indicating a perfect positive correlation. A perfect negative correlation means that as one variable increases, the other variable decreases at a constant rate. A perfect positive correlation means that as one variable increases, the other variable increases at a constant rate.

Pearson's R assumes that the variables being analysed are normally distributed, linearly related, and have equal variances. It is sensitive to outliers and can be affected by the scale of the variables being analyzed. If the variables are not normally distributed, a non-parametric correlation coefficient, such as Spearman's rank correlation coefficient, can be used instead.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is a preprocessing technique used in data analysis to transform variables to a common scale. It involves transforming variables so that they have a similar range of values. Scaling is performed to ensure that no single variable dominates the analysis and to improve the accuracy and efficiency of the analysis.

There are two common types of scaling: normalized scaling and standardized scaling.

Normalized scaling involves scaling the data so that it has a range of 0 to 1. Normalized scaling preserves the shape of the original data, but it can be affected by outliers and is sensitive to changes in the data.

Standardized scaling involves scaling the data so that it has a mean of 0 and a standard deviation of 1. Standardized scaling transforms the data to have a standard normal distribution, which makes it easier to compare and interpret the data. Standardized scaling is less affected by outliers and is more robust to changes in the data.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

The Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in a regression analysis. It is calculated for each independent variable in a regression model and is a measure of how much of the variance of this variable is predicted by other independent variables.

Sometimes, the value of VIF can be infinite. This happens when there is perfect multicollinearity in the regression model, which means that an independent variable in the model can be perfectly predicted by a linear combination of the other independent variables in the model. In this case, the VIF for the variable that can be perfectly predicted is undefined, resulting in an infinite VIF.

For example, if a regression model includes both a variable for temperature in Celsius and a variable for temperature in Fahrenheit, these two variables are perfectly collinear because they provide the same information. Another example of perfect multicollinearity is when a variable is defined as the sum or difference of two or more other variables in the model.

In order to address the issue of perfect multicollinearity and infinite VIF, it is necessary to remove one or more of the collinear variables from the model.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a sample of data to a theoretical probability distribution. It is a type of probability plot that plots the quantiles of the data against the quantiles of a theoretical distribution, typically the standard normal distribution.

In linear regression, the Q-Q plot is used to assess whether the residuals are normally distributed or not. If the residuals are normally distributed, the Q-Q plot of the residuals will form a straight line. If the residuals are not normally distributed, the Q-Q plot of the residuals will deviate from a straight line.

The Q-Q plot is an important tool in linear regression if the residuals are not normally distributed, this can affect the validity of the model. Additionally, non-normality in the residuals can indicate the presence of outliers or other problems with the model.

Submitted by:

Sharad Choudhury