Linear Regression

Subjective Questions and Answers

Sanjeev S

8/30/22

Master of Science in Machine
Learning & Artificial Intelligence

Contents

1. Assignment-based Subjective Questions

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

Analysis of various categorical variables was done using box plot and bar plot. Inference made from visualization of various plots is mentioned below:

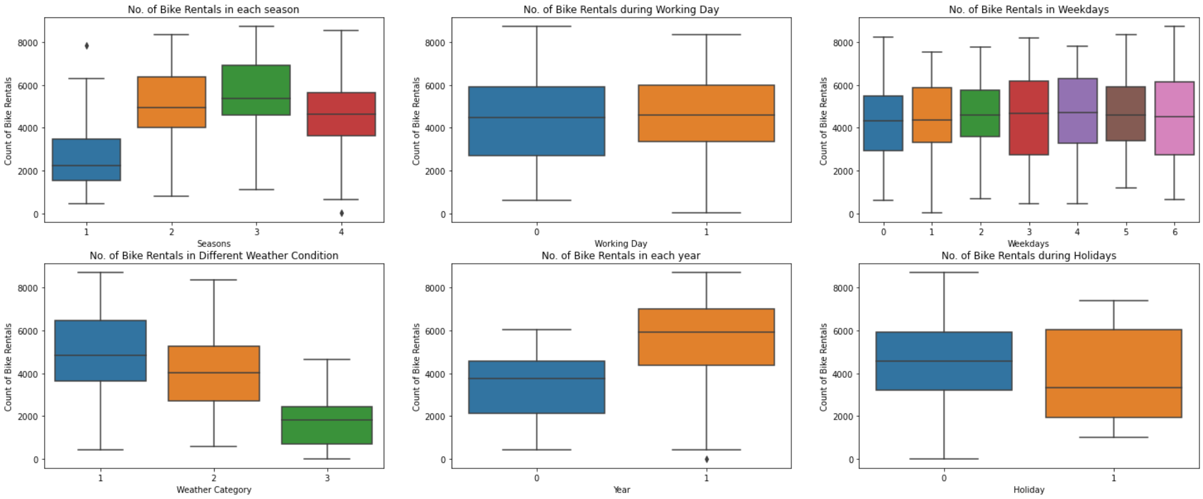


Figure 1: Box Plot of Various Categorical Features

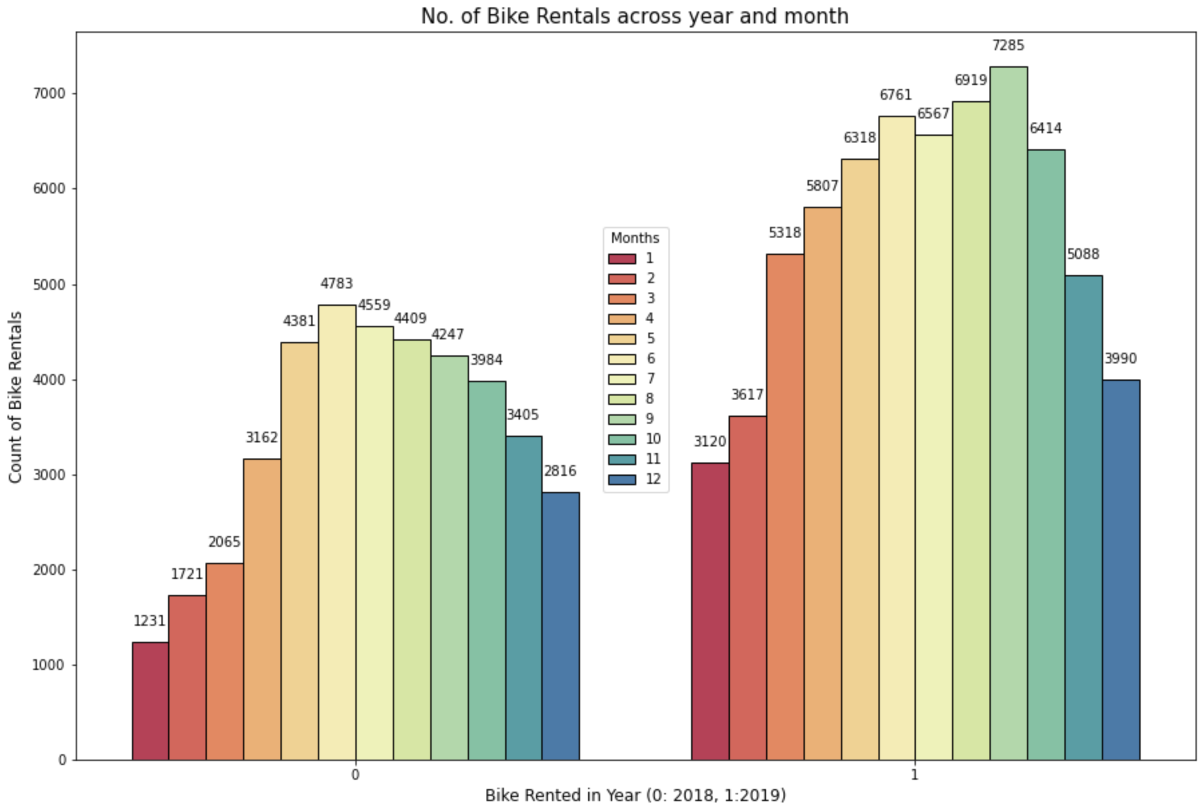


Figure 2: Bikes Rented in Year vs Months

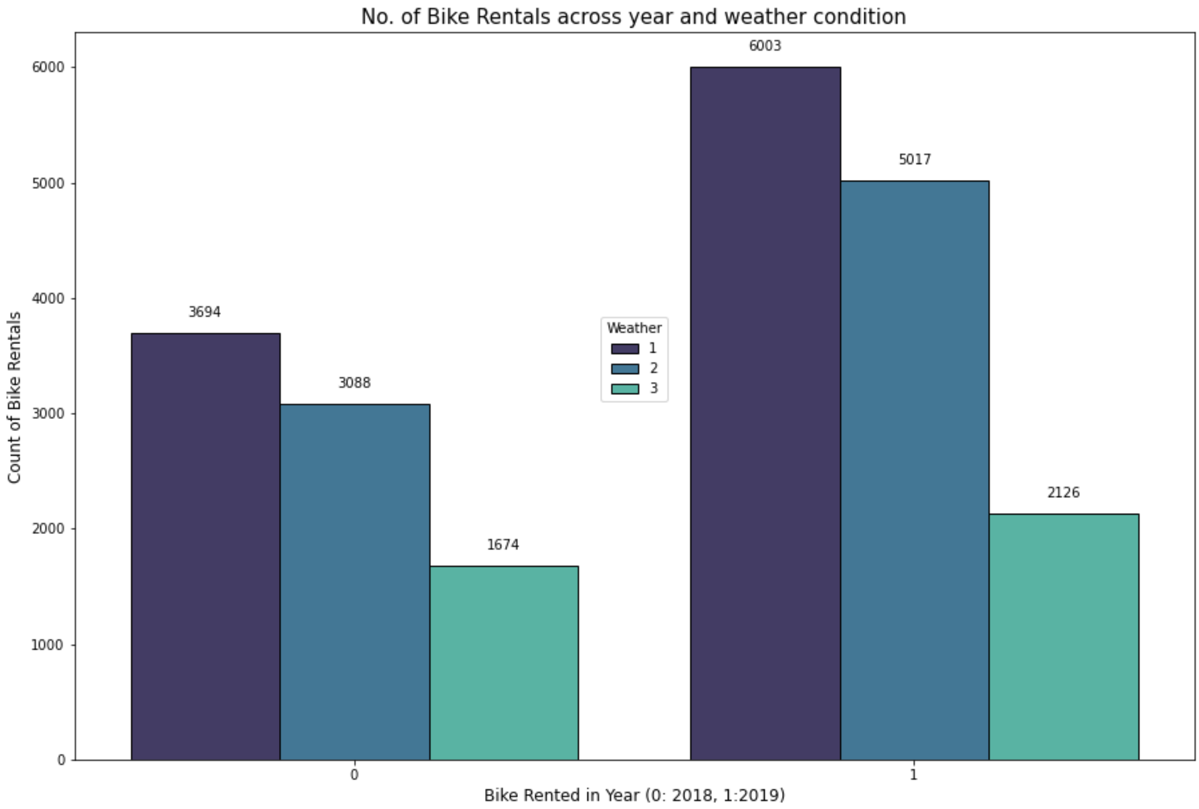


Figure 3: Bikes Rented in Year vs Weather

1) From box plot - 'No. of Bike Rentals in each season', judging by median more bikes are rented during Fall (3 in graph) season.

2) From box plot - 'No. of Bike Rentals during Working Day', median is similar for working and non-working days. Since spread is more for non-working days, there is less chance for bike rentals.

3) From box plot - 'No. of Bike Rentals in Weekdays', overall median looks similar across all days. Monday (6), Tuesday (0) and Wednesday (1) have more bike users.

4) From box plot - 'No. of Bike Rentals in Different Weather Condition', more bikes are rented during Clear, Few clouds, Partly cloudy, Partly cloudy weather (1 in x-axis).

5) From box plot - 'No. of Bike Rentals in each year', we can see that year 2019 have highest peak in rented bikes (higher median than 2018). Hence, we can conclude that every year, number of bike rentals can go up.

6) From box plot - 'No. of Bike Rentals during Holidays', we can see the spread is more during holidays but the median is less when compared to non-holidays. This means that more bike rentals happen in non-holidays. This might be because, during holidays users prefer another mode for transportation.

7) In 2019, more bikes were rented during September time. This clearly relates to why users rented bike more in 'Fall' season.

8) Number of bike rentals increased from 2018 to 2019, especially when the weather is Clear, Few clouds, Partly cloudy, Partly cloudy.

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

drop_first=True helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Example - Let's say we have 3 types of levels (A, B, C) in a Categorical column 'ABC' and we want to create dummy variable for that column. The dummy variable will look like:

A	**B**	**C**	**Levels**
1	0	0	**A**
0	1	0	**B**
0	0	1	**C**

Table 1: Dummy Variables for 'ABC'

Here even if we remove first column from table, we can still represent A as '00'. Let's see this by removing the column 'A'.

B	**C**	**Levels**
0	0	**A**
1	0	**B**
0	1	**C**

Table 2: Dummy Variables for 'ABC' with drop_first = True

Hence, we don't require 3 dummy variables to represent 3 levels. In general, a variable (feature) with n levels can be represented by n-1 dummy variables.

This same concept is used by drop_first feature of pandas.get_dummies.

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

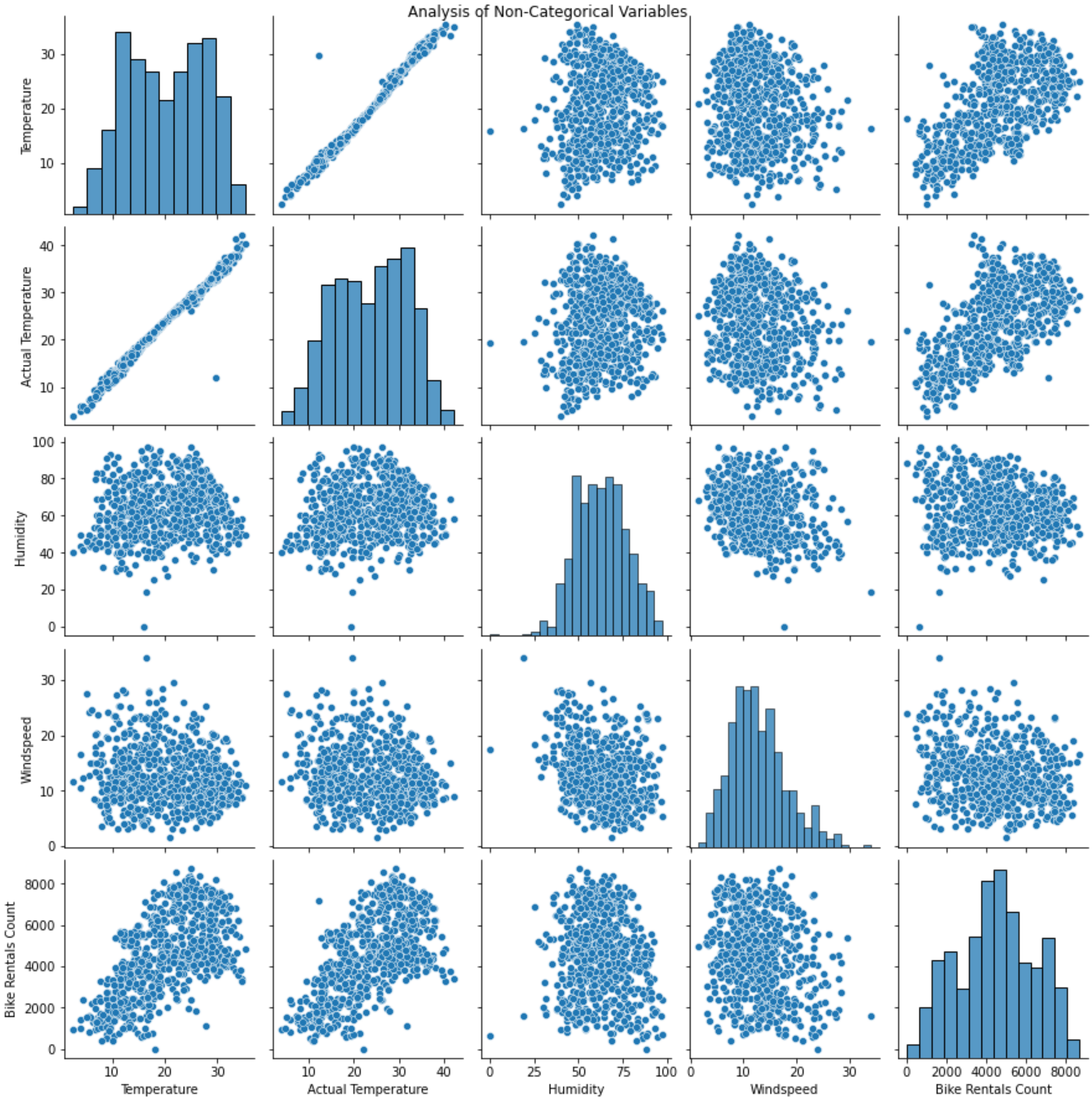


Figure 4: Pair Plot of Numerical Variables

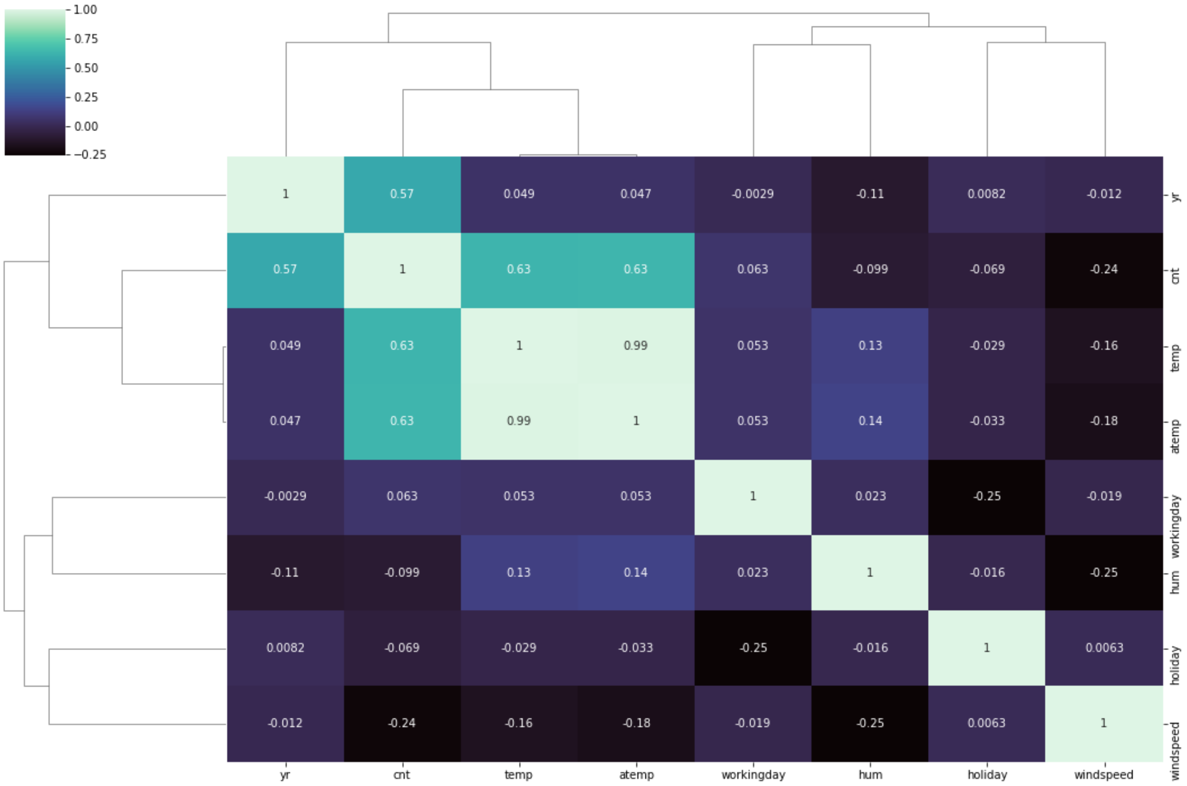


Figure 5: Correlation Mapping

Both ‘temp’ and ‘atemp’ has the highest correlation coefficient of 0.63 with target variable. After Recursive Feature Elimination, variable ‘atemp’ was removed as it was closely related to ‘temp’ and other variables.

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

Validated assumptions of Linear Regression using following 5 key assumptions:

1) Error Terms are normally distributed (Figure 6 – left).

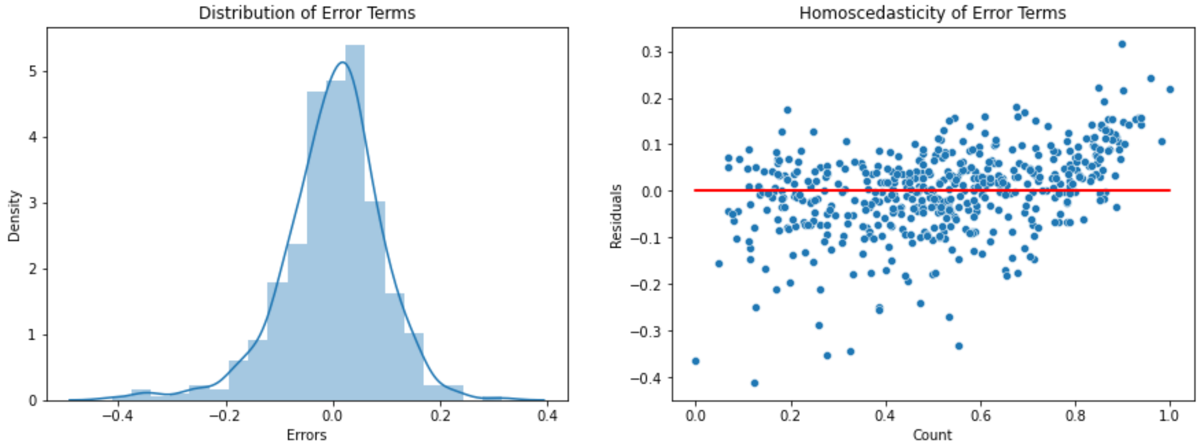


Figure 6: Distribution and Homoscedasticity of Error Terms

2) Error Terms have constant variance (homoscedasticity) (Figure 6 – right).

3) Linear Relationship between predictor variables and target variable (Figure 7).

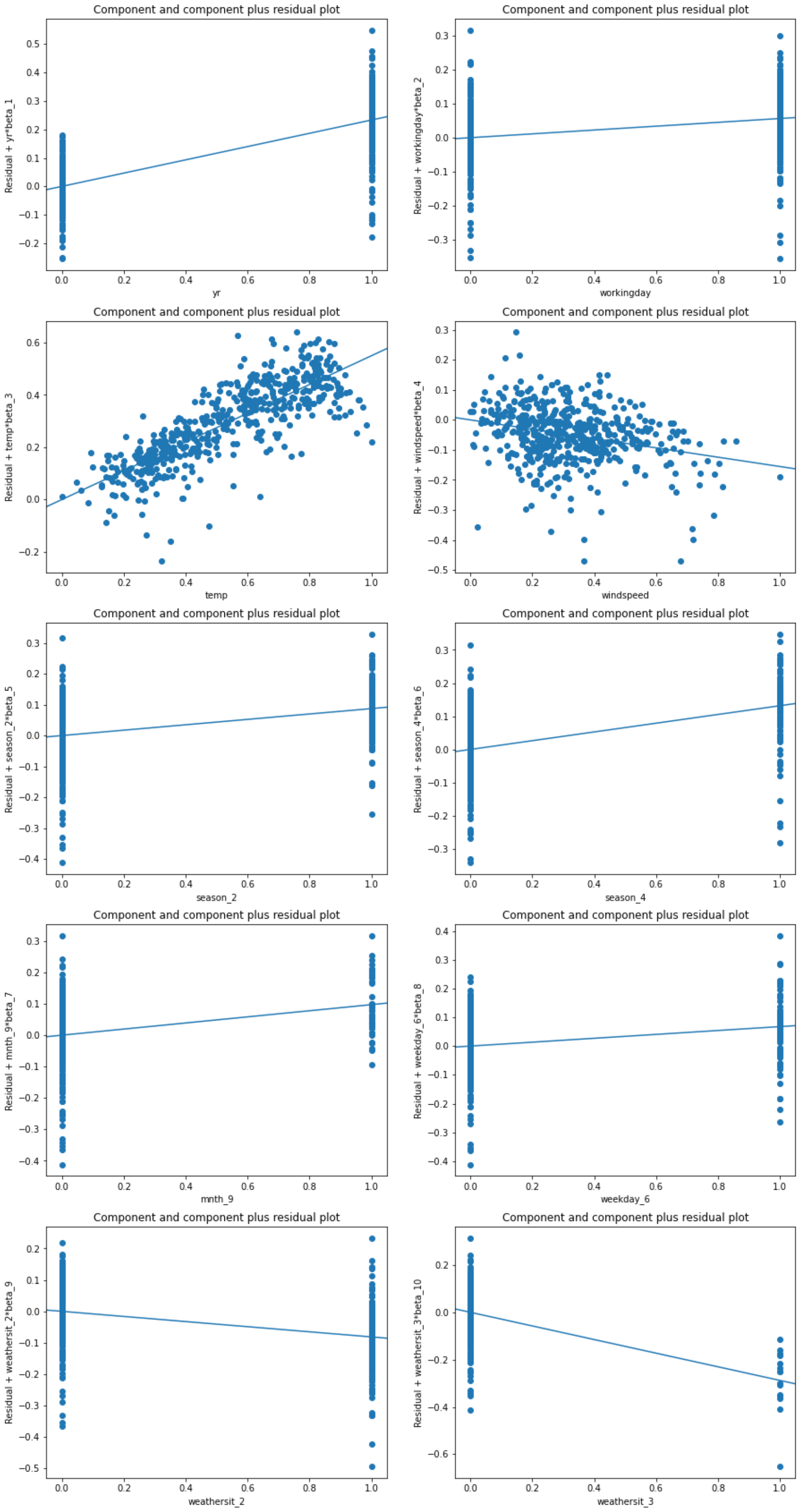


Figure 7: CCPR plot for linear relationship

4) No or Little multicollinearity. All Variance Inflation Factors (VIFs) of features are less than 5.

Features	VIF
temp	4.76
workingday	4.04
windspeed	3.43
yr	2.02
weekday_6	1.69
season_2	1.57
weathersit_2	1.53
season_4	1.4
mnth_9	1.2
weathersit_3	1.08

Table 3: VIF of various Features

5) No Auto-Correlation. Durbin-Watson value is **2.080** (acceptable range 1.50 - 2.50), so there is no autocorrelation in residuals.

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

The Following are the top 3 features contributing significantly towards explaining the demands of the shared bikes:

- **Temperature (temp)**: coefficient value of 0.55. As temperature variable increases bike rentals increases by 0.55 units, provided all other predictor variables are constant.
- **Year (yr)**: coefficient value of 0.23. Bike rentals can increase by 0.23 units / year, provided all other predictor variables are constant.
- **Weather Situation 3 (weathersit_3)**: coefficient value of -0.29. When weather is either Light Snow or Light Rain + Thunderstorm + Scattered clouds or Light Rain + Scattered clouds, bike rentals can decrease by 0.29 units, provided all other predictor variables are constant.

2. General Subjective Questions

2.1 Explain the linear regression algorithm in detail.

Linear Regression is a machine learning algorithm which is based on **supervised learning** category. It finds a best straight-line fitting on any given data, to find the best linear relationship between independent (Predictor) variables and dependent (Target) variable.

Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).
Mathematically the relationship can be represented with the help of following equation −
Y = mX + c
Here, Y is the dependent variable we are trying to predict.
X is the independent variable we are using to make predictions.
m is the slope of the regression line which represents the effect X has on Y
c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

Linear relationship can be positive or negative in nature:

- Positive relationship – when independent variable increases, dependent variable also increases.
- Negative relationship – when independent variable increases, dependent variable also decreases.

Linear regression is of the following two types:

- Simple Linear Regression – Only one independent variable.
- Multiple Linear Regression – Multiple independent variables.

5 key Assumptions of Linear Regression

- Relationship between variables – Linear regression model assumes that the relationship between target and feature variables must be linear.
- Normality of error terms – Error terms should be normally distributed.
- Auto-correlation – Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.
- Homoscedasticity – There should be no visible pattern in residual values and should have constant variance.
- Multi-collinearity – Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.

2.2 Explain the Anscombe’s quartet in detail.

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when plotted. It was developed by statistician **Francis Anscombe**.

The quartet is often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

The datasets are as follows. The x values are the same for the first three datasets.

Anscombe's quartet							
I		II		III		IV	
x	y	*x*	y	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

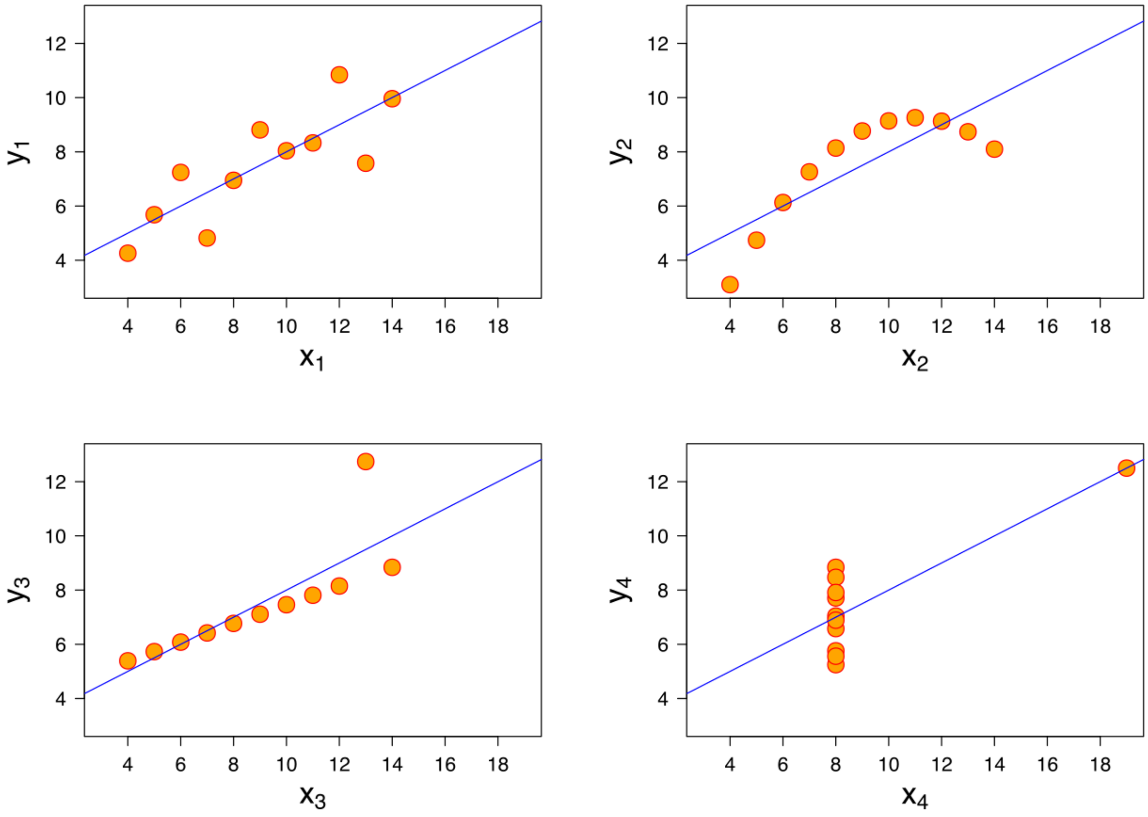
Table 4: Anscombe's Quartet Table

For all four datasets:

Property	**Value**	**Accuracy**
Mean of x	9	exact
Sample variance of x (s2x)	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y (s2y)	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	y = 3.00 + 0.500x	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression (R^2)	0.67	to 2 decimal places

Table 5: Statistical Figure of Anscombe's Quartet Dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



- Dataset I appear to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

In a nutshell, it is a better practice to visualize data and remove outliers before analysing it.

2.3 What is Pearson's R?

In statistics, the Pearson correlation coefficient — also known as **Pearson's R** is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations.

It is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

- A value of 1 means a total positive linear correlation. It means that if one variable increase then other will also increase
- A value of 0 means no correlation
- A value of -1 means a total negative correlation. It means that if one variable increase then other will decrease

There are several types of correlation coefficient formulas.

One of the most commonly used formulas is Pearson’s correlation coefficient formula.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

where:

- n is sample size
- xi, yi are the individual sample points indexed with i

Statistical inference based on Pearson's correlation coefficient often focuses on one of the following two aims:

- One aim is to test the null hypothesis that the true correlation coefficient ρ is equal to 0, based on the value of the sample correlation coefficient r.
- The other aim is to derive a confidence interval that, on repeated sampling, has a given probability of containing ρ.

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

Scaling is the process to normalize the data within a particular range. Scaling is a pre-processing step in linear regression analysis.

We scale a variable to make the computation of gradient descent faster. The step size of gradient descent is generally low for accuracy, if the data has some small variables (values in the range of 0-1) and some big variables (values in the range of 0 -1000) than the time taken by the gradient descent algorithm will be huge.

The two most discussed scaling methods are **Normalization** and **Standardization**. Normalization typically scales the values into a range of [0,1]. Standardization typically scales data to have a mean of 0 and a standard deviation of 1 (unit variance).

Normalised Scaling	**Standardized scaling**
Called min max scaling, scales the variable such that the range is 0-1	Values are centred around mean with a unit standard deviation
Good for non- gaussian distribution	Good for gaussian distribution
Value id bounded between 0 and 1	Value is not bounded
Outliers are also scaled	Does not affect outliers

Formula of Normalized scaling:

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

Formula of Standardized scaling:

$$x = \frac{x - mean(x)}{sd(x)}$$

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

The value of VIF is calculated by the below formula:

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

Where, ‘i’ refers to the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

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

A Q-Q plot is a probability plot, a graphical method for comparing two probability distributions by plotting their quantiles against each other. Its purpose is to check if the two sets of data came from the same distribution. It is a visual check of data. If the data is from same source, then the plot will appear as a line.

The Q-Q plot is used to see if the points lie approximately on the line. If they don’t, it means, our residuals aren’t Gaussian (Normal) and thus, our errors are also not Gaussian.

Given two cumulative probability distribution functions F and G, with associated quantile functions F−1 and G−1 (the inverse function of the CDF is the quantile function), the Q–Q plot draws the q-th quantile of F against the q-th quantile of G for a range of values of q. Thus, the Q–Q plot is a parametric curve indexed over [0,1] with values in the real plane R2.

Importance of Q-Q plot: Below are the points:

- The sample sizes do not need to be equal.
- Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.
- The q-q plot can provide more insight into the nature of the difference than analytical methods.