# Assignment-based Subjective Questions

# Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

# Total Marks: 3 marks (Do not edit)

# Answer: <Your answer for Question 1 goes below this line> (Do not edit)

# Season ,weather, holiday, month, working days and weekdays for the categorical variable of the dataset, a box plot was used to visualise these, These variables influenced our dependent variable in the following ways

1. **Working day:** it had little effect on dependent variable.
2. **Week days:** Weekends have a significance increase in the bike rentals hiring compared to the weekdays.
3. **Month :** September had the most bike rentals and December had the fewer, this observation is comparable to the one made in the weathersit the weather in the December is typically cold and snowy
4. **Holiday** rental were lower during the holidays
5. **Weather** when there is a heavy rain and snow there are no users indicating that the weather is extremely unfavourable, the highest count was observed when the weather forecast was clear or partly cloudy
6. **Season** the box plot revealed that a spring season had the lowest value of bike reantals while the fall season has the highest value of bike rentals ,summer and winter season had count values that one in the between.

# 

**Question 2.** Why is it important to use **drop\_first=True** during dummy variable creation? (Do not edit)

**Total Marks:** 2 marks (Do not edit)

# Answer: <Your answer for Question 2 goes below this line> (Do not edit)

# It is important in order to achieve k-1 dummy variables as it can be used to delete extra column while creating dummy variables. •For Example: We have three variables: Furnished, Semi-furnished and un-furnished. We can only take 2 variables as furnished will be 1-0, semi-furnished will be 0-1, so we don’t need unfurnished as we know 0-0 will indicate un-furnished. So we can remove it •It is also used to reduce the collinearity between dummy variables

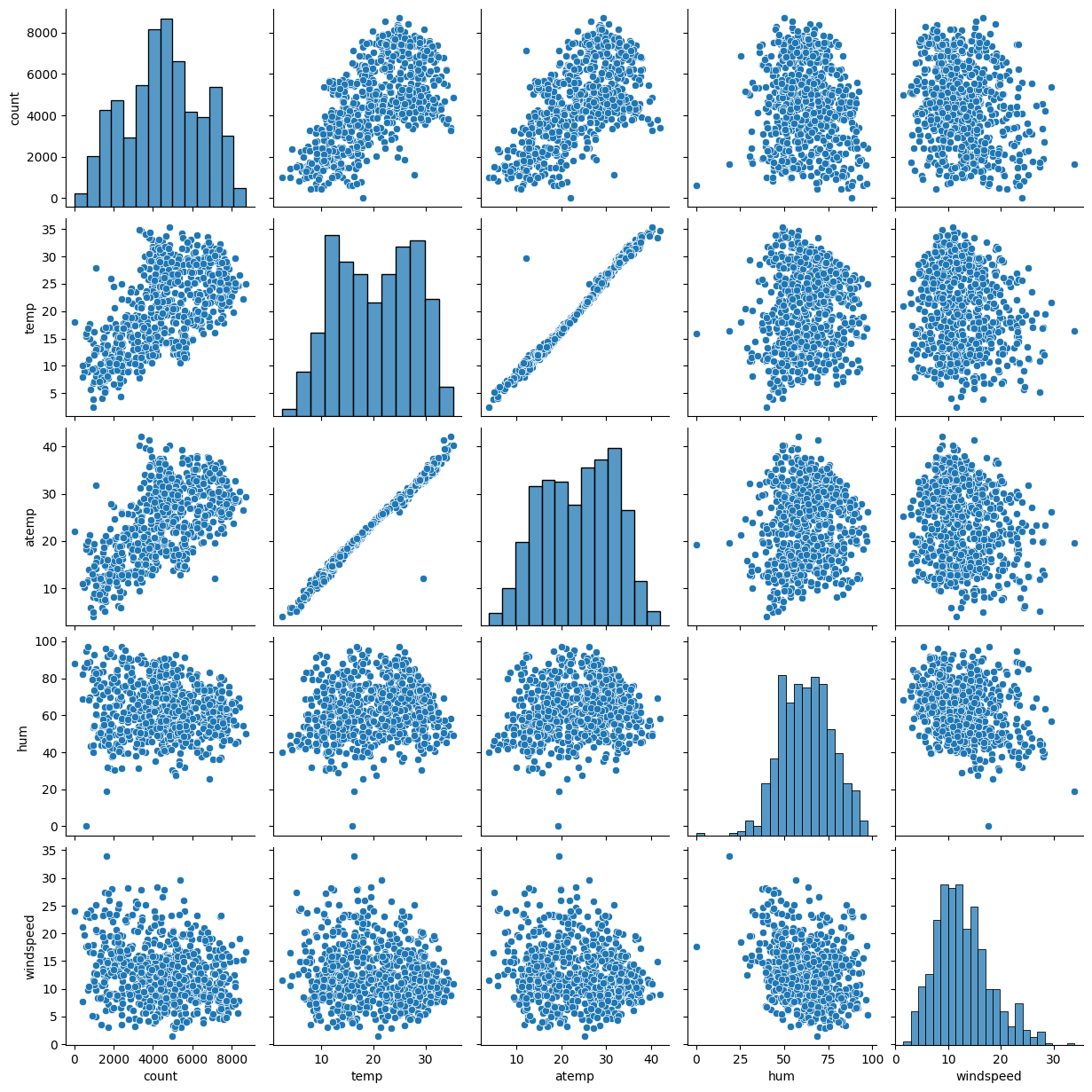
# If we do not use drop\_first = True, then n dummy variables will be created, and these predictors(n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap.

# Your dummy variable will be correlated if you don't remove the first column redundant. This may have a negative impact on some models and the effect is amplified when the cardinality is low Iterative models,for example may have difficulty convergent and list of variable important maybe distorted under argument is that having all dummy variables result in multicollinearity between them. We lose one column to keep everything under control.

**Question 3.** Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

**Total Marks:** 1 mark (Do not edit)

# Answer: <Your answer for Question 3 goes below this line> (Do not edit)

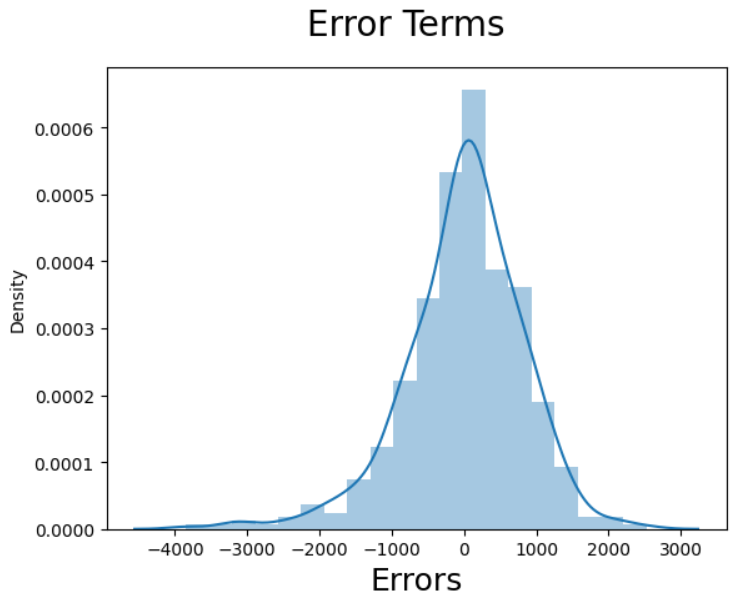


atemp and temp both have same correlation with target variable of 0.63 which is the highest among all numerical variables.

**Question 4.** How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

# Answer: <Your answer for Question 4 goes below this line> (Do not edit)



The distribution of residuals should be normal and centred around 0 because the mean is zero. We test this residuals assumption by producing a distplot of residuals to see if the following a normal distribution or not. The residuals are scatter round mean is equal to zero as the in the above diagram

**Question 5.** Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)

**Total Marks:** 2 marks (Do not edit)

# Answer: <Your answer for Question 5 goes below this line> (Do not edit)

**Temp**: with a coefficient of 0.3917, a unit increase in the temp variable increases the number of bike rentals by 0.3917units.

**Year:** With a coefficient of 0.2001, a unit increase in the yr variable increase the number of bike rentals by 0.2001.

**Weathersit\_bad**: With a coefficient of (-0.2305), a unit increase in the weather\_bad variable reduces the number of bike rentals by 0.2305 as compared to weathershit\_good.

# General Subjective Questions

**Question 6.** Explain the linear regression algorithm in detail. (Do not edit)

**Total Marks:** 4 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

Linear regression is a machine learning algorithm that uses a linear equation to model the relationship between a dependent variable and one or more independent variables. It's a basic form of data analysis that predicts unknown values by using known, related data

Linear regression finds the best-fit line or surface that minimizes the difference between predicted and actual values. It does this by estimating the coefficients of the linear equation that best predicts the dependent variable.

**Types of linear regression**

A model with one independent variable is a simple linear regression, while a model with more than one is a multiple linear regression.

**How to use linear regression**

For example, you can use linear regression to predict future expenses based on your past income and expenses. You can also use it to estimate the number of customers you might get for a sales promotion by looking at the number of customers you got for previous promotions.

**Simple linear regression**

Simple linear regression is defined by the linear function:

**Y= β0\*X + β1 + ε**

β0 and β1 are two unknown constants representing the regression slope, whereas ε (epsilon) is the error term.

You can use simple linear regression to model the relationship between two variables, such as these:

Rainfall and crop yield

Age and height in children

Temperature and expansion of the metal mercury in a thermometer

**Multiple linear regression**

In multiple linear regression analysis, the dataset contains one dependent variable and multiple independent variables. The linear regression line function changes to include more factors as follows:

**Y= β0\*X0 + β1X1 + β2X2+…… βnXn+ ε**

As the number of predictor variables increases, the β constants also increase correspondingly.

 Multiple linear regression models multiple variables and their impact on an outcome:

Rainfall, temperature, and fertilizer use on crop yield

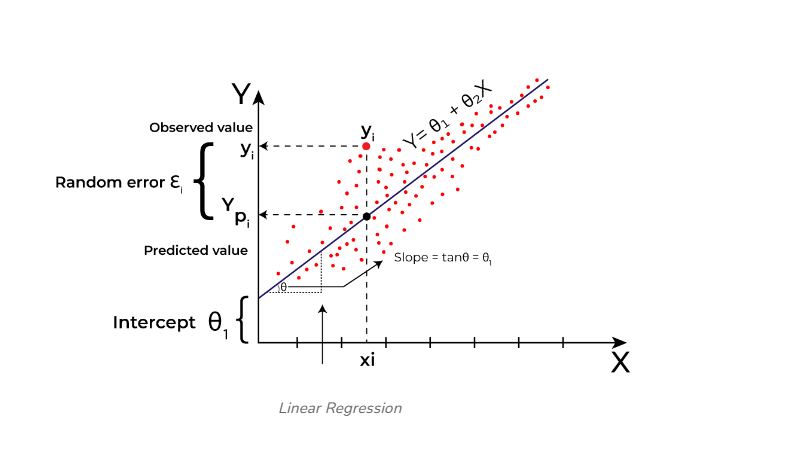
Diet and exercise on heart disease

Wage growth and inflation on home loan rates

**Linear regression** is a type of [supervised machine learning](https://www.geeksforgeeks.org/supervised-machine-learning/) algorithm that computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to observed data.

When there is only one independent feature, it is known as **[Simple Linear Regression](https://www.geeksforgeeks.org/simple-linear-regression-using-r/" \t "https://www.geeksforgeeks.org/ml-linear-regression/_blank)**, and when there are more than one feature, it is known as **[Multiple Linear Regressio](https://www.geeksforgeeks.org/ml-multiple-linear-regression-using-python/" \t "https://www.geeksforgeeks.org/ml-linear-regression/_blank)**[n](https://www.geeksforgeeks.org/ml-multiple-linear-regression-using-python/" \t "https://www.geeksforgeeks.org/ml-linear-regression/_blank).

Similarly, when there is only one dependent variable, it is considered **[Univariate Linear Regression](https://www.geeksforgeeks.org/univariate-linear-regression-in-python/" \t "https://www.geeksforgeeks.org/ml-linear-regression/_blank)**, while when there are more than one dependent variables, it is known as **[Multivariate Regression](https://www.geeksforgeeks.org/multivariate-regression/" \t "https://www.geeksforgeeks.org/ml-linear-regression/_blank)**.



**Question 7.** Explain the Anscombe’s quartet in detail. (Do not edit)

**Total Marks:** 3 marks (Do not edit)

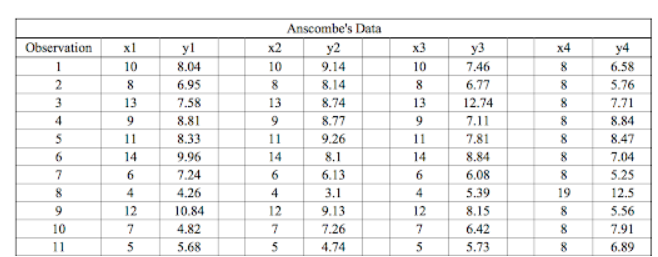
**Answer:** Please write your answer below this line. (Do not edit)

# Anscombe’s quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

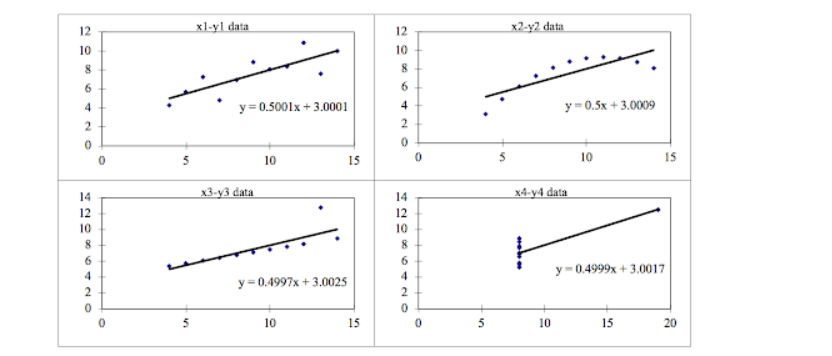
Anscombe’s quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics.  It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

Anscombe’s quartet tells us about the importance of [visualizing data](https://builtin.com/learn/data-visualization" \t "https://builtin.com/data-science/_blank) before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

We can define these four plots as follows:



However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn’t interpretable by any regression algorithm, as you can see below:



Anscombe’s Quartet Four Datasets

Data Set 1: fits the linear regression model pretty well.

Data Set 2: cannot fit the linear regression model because the data is non-linear.

Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.

Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

**Question 8.** What is Pearson’s R? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

The Pearson correlation coefficient (r) is the most widely used correlation coefficient and is known by many names:

Pearson’s R

Bivariate correlation

Pearson product-moment correlation coefficient (PPMCC)

The correlation coefficient

The Pearson correlation coefficient is a [descriptive statistic](https://www.scribbr.com/statistics/descriptive-statistics/), meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

Although interpretations of the relationship strength (also known as [effect size](https://www.scribbr.com/statistics/effect-size/)) vary between disciplines, the table below gives general rules of thumb:

| **Pearson correlation coefficient (r) value** | **Strength** | **Direction** |
| --- | --- | --- |
| Greater than .5 | Strong | Positive |
| Between .3 and .5 | Moderate | Positive |
| Between 0 and .3 | Weak | Positive |
| 0 | None | None |
| Between 0 and –.3 | Weak | Negative |
| Between –.3 and –.5 | Moderate | Negative |
| Less than –.5 | Strong | Negative |

The Pearson correlation coefficient is also an [inferential statistic](https://www.scribbr.com/statistics/inferential-statistics/), meaning that it can be used to [test statistical hypotheses](https://www.scribbr.com/statistics/hypothesis-testing/). Specifically, we can test whether there is a significant relationship between two variables.

**Question 9.** What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)

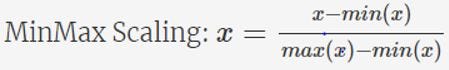
**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

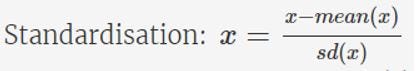
# **What?** It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

**Why?**Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

# **Normalization/Min-Max Scaling:**It brings all of the data in the range of 0 and sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.



Standardization Scaling: Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).



1. sklearn.preprocessing.scale helps to implement standardization in python.
2. One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

**Question 10.** You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

The variance inflation factor (VIF) can be infinite when there is perfect multicollinearity between variables, or when one independent variable is a linear combination of the other independent variables:

**Perfect multicollinearity**

When the regressor is equal to a linear combination of other regressors, the VIF tends to infinity.

**Linear combination**

When one independent variable can be exactly written as a linear combination of the others, the corresponding VIF will be infinite.

**R² value of 1**

When calculating the VIF for one independent variable using all the other independent variables, if the R² value comes out to be 1, the VIF will become infinite.

**Question 11.** What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

# This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

# Few advantages:

# a) It can be used with sample sizes also

# b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

# It is used to check following scenarios:

# If two data sets —

# i. come from populations with a common distribution

# ii. have common location and scale

# iii. have similar distributional shapes

# iv. have similar tail behavior