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?**

   The categorical variable used in the dataset: season , yr(year) , holiday, weekday ,workingday, and weathersit(weather situation) and mnth(month) . These were visualized using a boxplot.

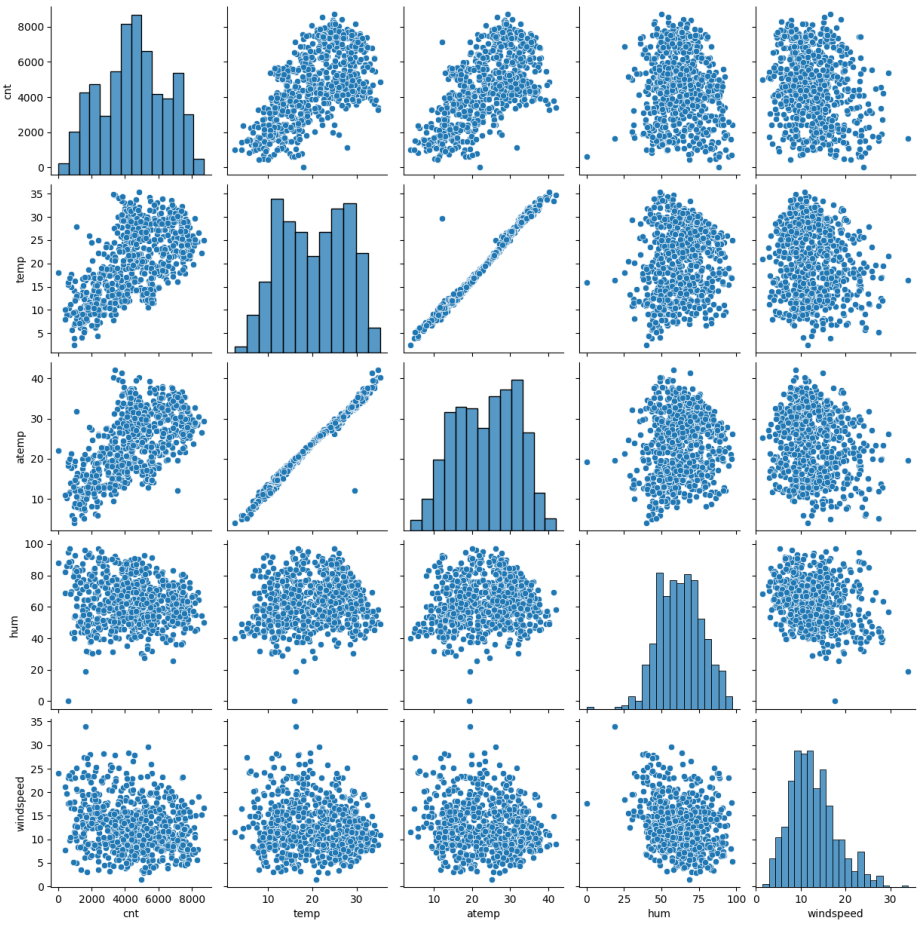
   These variables had the following effect on our dependant variable: -

- Season - For the variable season, we can clearly see that the category 3: Fall, has the highest median, which shows that the demand was high during this season. It is least for 1: spring.
- Yr - The year 2019 had a higher count of users as compared to the year 2018.
- Holiday - rentals reduced during holiday.
- Weekday - The bike demand is almost constant throughout the week.
- Workingday – From the "Workingday" boxplot we can see those maximum bookings happening between 4000 and 6000, that is the median count of users is constant almost throughout the week. There is not much of difference in booking whether its working day or not.
- Weathersit - There are no users when there is heavy rain/ snow indicating that this weather is quite adverse. Highest count was seen when the weather situation was Clear, Partly Cloudy.
- Mnth - The number of rentals peaked in September, whereas they peaked in December. This observation is consistent with the observations made regarding the weather. As a result of the typical substantial snowfall in December, rentals may have declined

2. **Why is it important to use `drop_first=True` during dummy variable creation?**
   drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

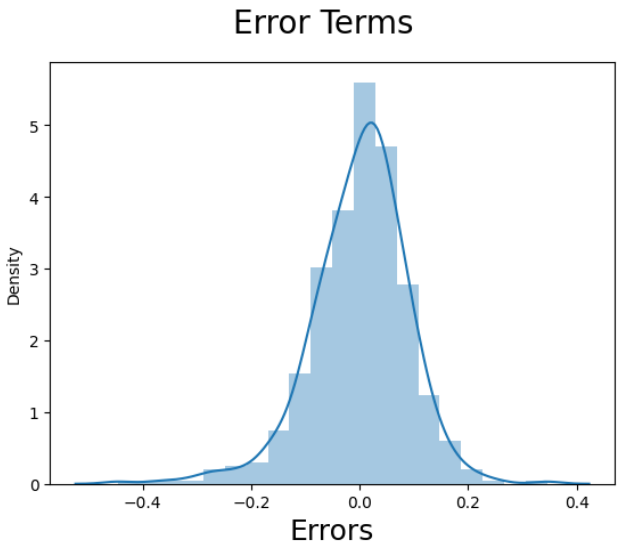
   Consider a Categorical column with 3 types of values, we want to create dummy variable for that column. If one variable is neither furnished nor semi_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   “temp” and “atemp” are the two numerical variables which are highly correlated with the target variable (cnt).


cnt
temp
atemp
hum
windspeed
0
2000
4000
6000
8000
5
10
15
20
25
30
35
40
0
20
40
60
80
100
0
5
10
15
20
25
30
0
2000
4000
6000
8000
10
20
30
10
20
30
40
0
25
50
75
100
0
10
20
30

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   We have done following tests to validate assumptions of Linear Regression:

   a. There should be linear relationship between independent and dependent variables. We visualised the numeric variables using a pairplot to see if the variables are linearly related or not. (ref. see above question's pairplot)
   b. Residuals distribution should follow normal distribution and centred around 0 (mean = 0). We validated this assumption about residuals by plotting a distplot of residuals and saw if residuals are following normal distribution or not.



   c. linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other. We calculated the VIF (Variance Inflation Factor) to quantify how strongly the feature variables in the new model are associated with one another. For more information, consult the notebook.

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 significant features are:
   1. temp - coefficient : 0.437655
   2. yr - coefficient : 0.234287
   3. weathersit_Light Snow & Rain - coefficient : -0.292892

General Subjective Questions

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

Linear regression is a statistical model that examines the linear association between a dependent variable and a given set of independent variables. This model indicates that when the value of one or more independent variables changes (either increases or decreases), the value of the dependent variable will also change correspondingly (either increase or decrease).

Mathematical formula : $\mathbf{Y = \beta_o + \beta_1 X}$

where

Y is the dependent variable.

X is the independent variable.

β1 is the slope of the regression line which represents the e7ect X has on Y

β_o is a constant, known as the Y-intercept. If X = 0, Y would be equal to β_o

There are two types of Linear Regression

a. Simple Linear Regression

b. Multiple Linear Regression

Linear Regression Model has following assumptions.

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.

Autocorrelation

The Linear regression model also assumes that there is minimal or no autocorrelation present in the data. Autocorrelation refers to the existence of a relationship between residual errors.

Relationship between variables

There should be linear relationship among variables.

Normality of error terms

Error terms should be normally distributed.

Homoscedasticity

There should be no visible pattern in residual values.

2. **Explain the Anscombe's quartet in detail.**

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet they di7er significantly when graphically represented and analysed. This quartet was created by the statistician Francis Anscombe in 1973 to emphasize the importance of visualizing data and not relying solely on summary statistics.

The four datasets in Anscombe's quartet share the same mean, variance, correlation coe7icient, and linear regression line, but they exhibit distinct patterns when graphed. Each dataset consists of 11 (x, y) pairs. Let's take a closer look at each of the four sets:

Dataset I:

- **Description:** A simple linear relationship.
- **Graphical Representation:** A scatter plot with a clear linear trend.
- **Summary Statistics:**

• Mean of x: 9.0
• Mean of y: 7.5
• Variance of x: 11.0
• Variance of y: 4.12
• Correlation coe7icient: 0.82
• Linear regression: y = 3.0 + 0.5 * x

Dataset II:

• **Description:** A non-linear relationship.
• **Graphical Representation:** A scatter plot with a curve; not a simple linear relationship.
• **Summary Statistics:**
• Mean of x: 9.0
• Mean of y: 7.5
• Variance of x: 11.0
• Variance of y: 4.12
• Correlation coe7icient: 0.82
• Linear regression: y = 3.0 + 0.5 * x

Dataset III:

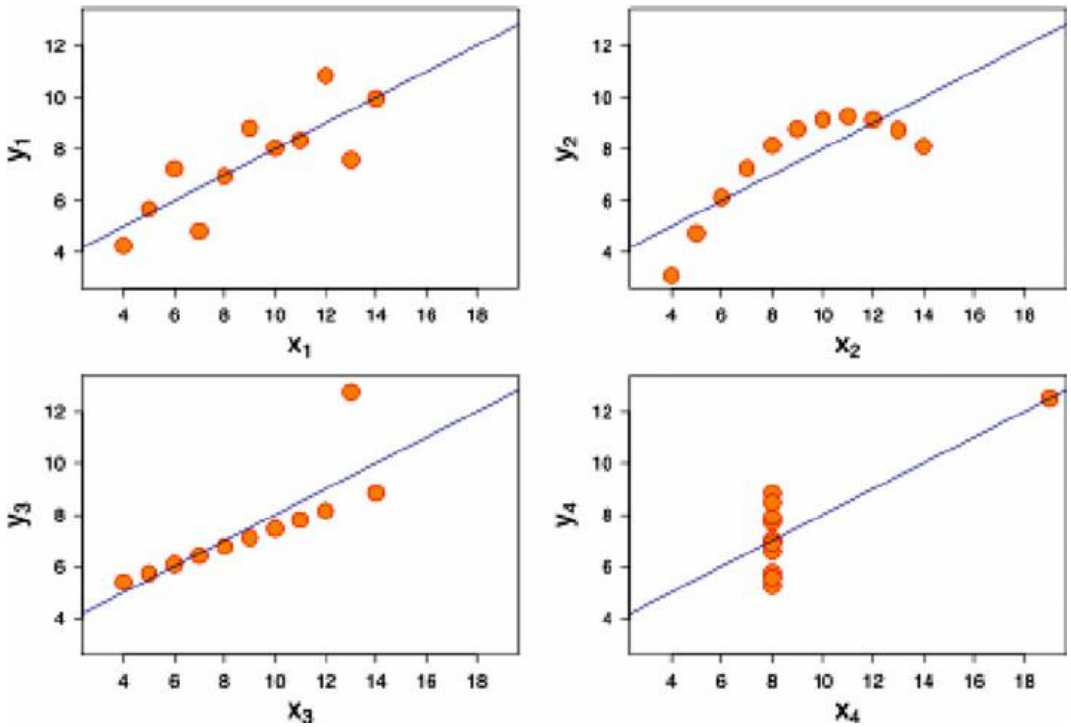
• **Description:** A situation with an outlier.
• **Graphical Representation:** A scatter plot with a clear linear trend, except for one outlier.
• **Summary Statistics:**
• Mean of x: 9.0
• Mean of y: 7.5
• Variance of x: 11.0
• Variance of y: 4.12
• Correlation coe7icient: 0.82
• Linear regression: y = 3.0 + 0.5 * x

Dataset IV:

• **Description:** Two distinct clusters.
• **Graphical Representation:** Two separate clusters with di7erent linear relationships.
• **Summary Statistics:**
• Mean of x: 9.0
• Mean of y: 7.5
• Variance of x: 11.0
• Variance of y: 4.12
• Correlation coe7icient: 0.82
• Linear regression: y = 3.0 + 0.5 * x

When we graph these four datasets on an x/y coordinate plane, it becomes evident that they exhibit identical regression lines. However, each dataset conveys a distinct narrative.

The quartet illustrates the importance of visually inspecting data through plots and graphs, as relying solely on summary statistics can be misleading. Even though these datasets share similar basic statistics their underlying structures are vastly di7erent. This emphasizes the need for exploratory data analysis and visualization to gain a more comprehensive understanding of the data.



The quartet illustrates the importance of visually inspecting data through plots and graphs, as relying solely on summary statistics can be misleading. Even though these datasets share similar basic statistics, their underlying structures are vastly di7erent. This emphasizes the need for exploratory data analysis and visualization to gain a more comprehensive understanding of the data.

3. What is Pearson's R?

Pearson's r is a numerical summary of the strength of the linear association between the variables. It value ranges between -1 to +1. It shows the linear relationship between two sets of data. In simple terms, it tells us "can we draw a line graph to represent the data?"

Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

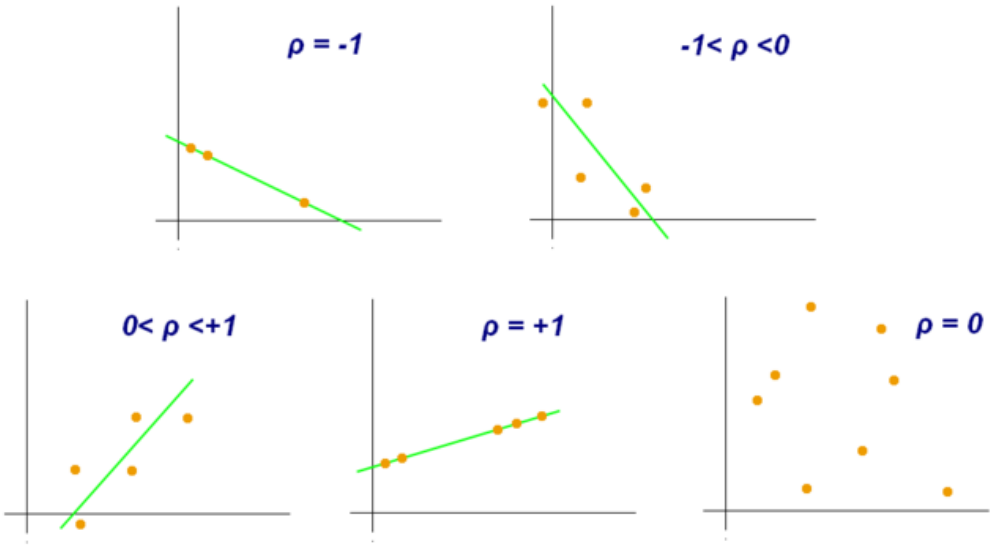
x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

As can be seen from the graph below, r = 1 means the data is perfectly linear with a positive slope r = -1 means the data is perfectly linear with a negative slope r = 0 means there is no linear association



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

Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data pre-processing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.

- Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbours and Neural Networks.
- Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

The VIF (Variance Inflation Factor) gives how much the variance of the coefficient estimate is being inflated by collinearity. If there is perfect correlation, then VIF = infinity. It gives a basic quantitative idea about how much the feature variables are correlated with each other. It is an extremely important parameter to test our linear model.

$$VIF = \frac{1}{1 - R^2}$$

Where R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables. If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1. So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity" The numerical value for VIF tells you (in decimal form) what percentage the variance (i.e. the standard error squared) is inflated for each coefficient. For example, a VIF of 1.9 tells you that the variance of a particular coefficient is 90% bigger than what you would expect if there was no multicollinearity — if there was no correlation with other predictors.
A rule of thumb for interpreting the variance inflation factor:

- 1 = not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.

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 plot of the quantiles of the first data set against the quantiles of the second data set. It is used to compare the shapes of distributions. The Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. The q-q plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behaviour?