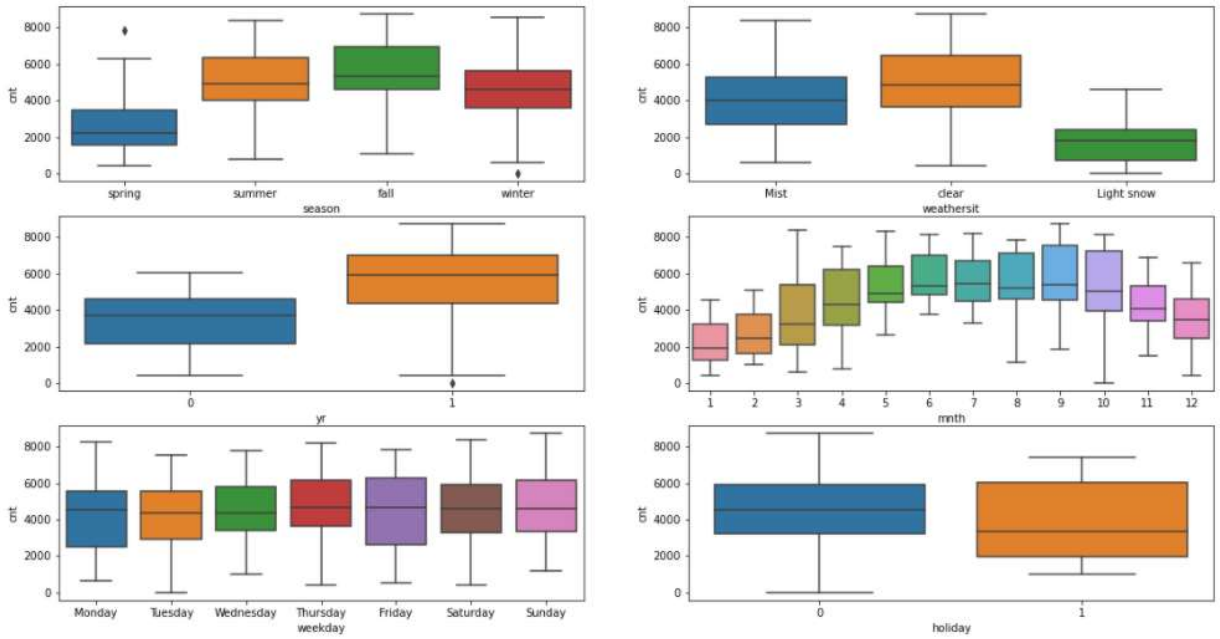
Assignment-based Subjective Questions

Q1.From your analysis of the categorical variables from the dataset, what could you infer about

their effect on the dependent variable?



- The demad of bike is less in the month of spring when compared with other seasons
- The demand bike increased in the year 2019 when compared with year 2018.
- Month Jun to Sep is the period when bike demand is high. The Month Jan is the lowest demand month.
- Bike demand is less in holidays in comparison to not being holiday.
- The demand of bike is almost similar throughout the weekdays.
- There is no significant change in bike demand with workign day and non working day.
- The bike demand is high when weather is clear and Few clouds however demand is less in case of Lightsnow and light rainfall. - We do not have any dat for Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog , so we can not derive any conclusion. May be the company is not operating on those days or there is no demand of bike.

Q2. 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 variable because more dummy features make it harder for the algorithm to fit or even worse make it easier to overfit.

Example:- Imagine you are looking at a coin flip, and have a feature called head, you do not need a column tail because you already know it via head=False. Same applies to other features like your month, if jan to nov are false it is clear that it is december.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation

with the target variable?

Both **" `atem` "** and **" `temp` "** have high correlation with traget vairbales.

Q4.How did you validate the assumptions of Linear Regression after building the model on the

training set?

There are four principal assumptions which justify the use of linear regression models for purposes of inference or prediction:

- linearity and additivity of the relationship between dependent and independent variables:

```
(a) The expected value of dependent variable is a straight-line func
tion of each independent variable, holding the others fixed.

(b) The slope of that line does not depend on the values of the othe
r variables.

(c)  The effects of different independent variables on the expected
 value of the dependent variable are additive.
```

- statistical independence of the errors (in particular, no correlation between consecutive errors in the case of time series data)
- homoscedasticity (constant variance) of the errors

```
(a) versus time (in the case of time series data)

(b) versus the predictions

(c) versus any independent variable
```

- normality of the error distribution.

Q5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes?

Top three features are :-

- 1.Temp
- 2.Yr
- 3.Winter Season

General Subjective Questions

Q1.Explain the linear regression algorithm in detail.

Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog). There are two main types:

Simple regression

Simple linear regression uses traditional slope-intercept form, where m and b are the variables our algorithm will try to "learn" to produce the most accurate predictions. x represents our input data and y represents our prediction.

y=mx+b

Multivariable regression

A more complex, multi-variable linear equation might look like this, where m represents the coefficients, or weights, our model will try to learn.

y=b+m1x1+m2x2+m3x3.....

The variables x1,x2,x3 represent the attributes, or distinct pieces of information, we have about each observation. For sales predictions, these attributes might include a company's advertising spend on radio, TV, and newspapers.

Sales=m1Radio+m2TV+m3News

Q2. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

Simple understanding:

Once Francis John “Frank” Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below.

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

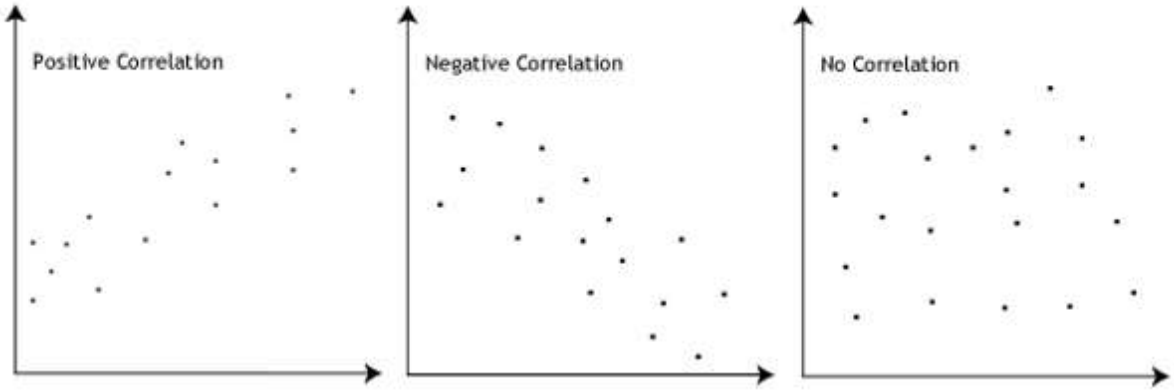
After that, the council analyzed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

Q3. What is Pearson’s R?

In statistics, the Pearson correlation coefficient (PCC), also referred to as **Pearson's R**, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction) r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions) r = 0 means there is no linear association r > 0 < 5 means there is a weak association r > 5 < 8 means there is a moderate association r > 8 means there is a strong association



Formula

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

Here,

- r=correlation coefficient
- xi=values of the x-variable in a sample
- bar^x=mean of the values of the x-variable
- yi=values of the y-variable in a sample
- bar^y=mean of the values of the y-variable

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

and standardized scaling?

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.

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 typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

S.NO.	Normalisation	Standardisation
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called `MinMaxScaler` for Normalization.	Scikit-Learn provides a transformer called `StandardScaler` for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization

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

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity.

In order to determine VIF, we fit a regression model between the independent variables. For example, we would fit the following models to estimate the coefficient of determination R1 and use this value to estimate the VIF:

$X_1=C+ m2 X_2+m3 X_3+\cdots$

$〖VIF〗_1=1/(1-R_1^2)$

Next, we fit the model between X2 and the other independent variables to estimate the coefficient of determination R2:

$X_2=C+ m1 X_1+m3 X_3+\cdots$

$〖VIF〗_2=1/(1-R_2^2)$

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity.

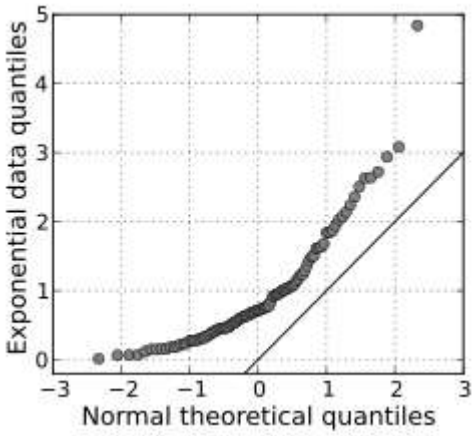
This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.