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? (3 marks)

Ans: The equation for the best fitted line is

cnt = 0.235 * yr - 0.086 * holiday + 0.475 * temp - 0.132 * windspeed - 0.103 * season_spring + 0.05 * season_winter - 0.256 * weathersit_L_Rain_Snow - 0.061 * mnth_Jul + 0.049 * mnth_Sep

In the above mentioned equation the following variables are the categorical variables.

- yr, holiday, season_spring, season_winter, weathersit_L_Rain_Snow , mnth_Jul & mnth_Sep

The yr , season_winter and mnth_Sep is having positive co-efficient I.e. it increases the value of the dependent variable cnt.

Whereas holiday, season_spring,weathersit_L_Rain_Snow & mnth_Jul is having negetive co-efficient I.e. it decreases the value of the dependent variable cnt.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Ans: While creating dummy variables **drop_first=True** commands removes the first column created in the new dummy data frame. Lets take an example of our season categorical variable.

The table for the dummy variable looks similar to the below table. (without **drop_first=True command)**

	Fall	**Spring**	**Winter**	**Summer**
0	**1**	**0**	**0**	**0**
…	**0**	**1**	**0**	**0**
…	**0**	**0**	**1**	**0**
…	**0**	**0**	**0**	**1**

Now, we don't need three columns. we can drop the `fall` column, as the type of seasons can be identified with just the last three columns where —

- 100 will correspond to `spring`

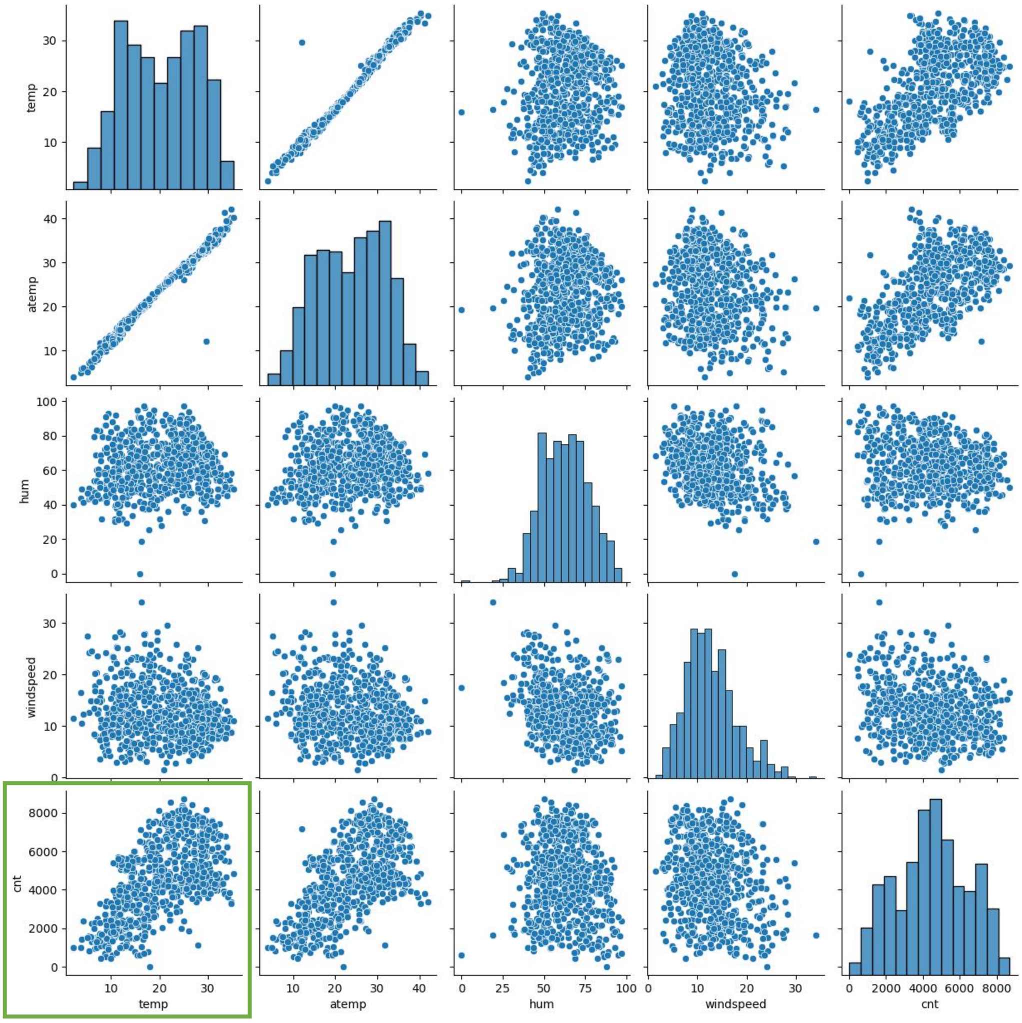
- 010 will correspond to `summer`

- 101 will correspond to `winter`

So any categorical variables with ‘n’ levels can be created with ‘n-1’ dummy variables. **Also it decreases the computation time.**

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans: The Below image shows the pair plot among the target variable ‘cnt ’ and all other numerical variables in the dataset.

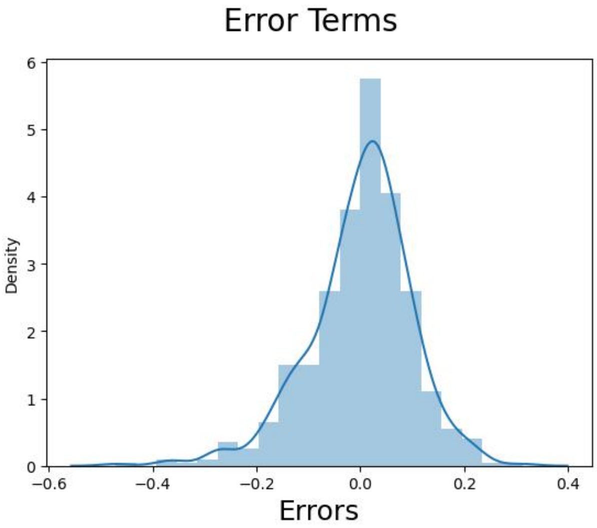


It is clearly visible that ‘temp’ is having the highest co-relation with the target variable.

Also ‘temp’ and ‘atemp’ is also having very high positive co-relation.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans: The major assumptions of Linear Regression is that the error terms has been normally distributed. The histogram of the error term shows that it has been normally distributed with mean zero.



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans: The equation for the best fitted line is

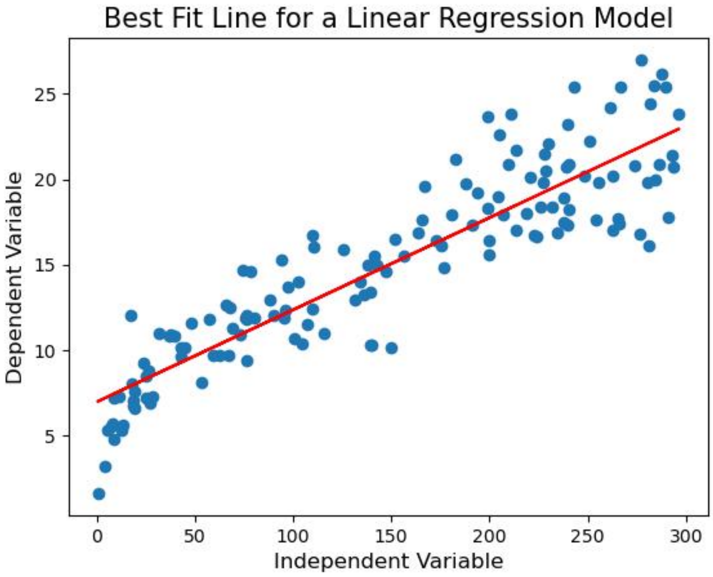
cnt = 0.235 * yr - 0.086 * holiday + 0.475 * temp - 0.132 * windspeed - 0.103 * season_spring + 0.05 * season_winter - 0.256 * weathersit_L_Rain_Snow - 0.061 * mnth_Jul + 0.049 * mnth_Sep

So, from the co-efficient value it is evident that "temp" hold the highest significance towards explaining the demand of the shared bikes. Followed by 'yr' and 'weathersit_L_Rain_Snow'.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans: "Linear regression" uses a linear relationship between an independent and dependent variable to forecast future events. It is a statistical technique for predictive analysis used in data science and machine learning.



A sloped straight line represents the linear regression model.

In the above figure,

X-axis = Independent variable

Y-axis = Output / dependent variable

Line of regression = Best fit line for a model

The predictor or explanatory variable that does not change as a result of changes in other variables is also known as the independent variable. But variations in the independent variable also affect the dependent variable. The response or result variable under analysis or study is known as the dependent variable, and the regression model predicts its value.

The equation of the best fit regression line $y = \beta_0 + \beta_1 X$

y = output variable. Variable y represents the continuous value that the model tries to predict.

x = input variable. In machine learning, x is the feature, while it is termed the independent variable in statistics. Variable x represents the input information provided to the model at any given time.

β_0 = y-axis intercept.

β_1 = the regression coefficient or scale factor.

2. Explain the Anscombe's quartet in detail. (3 marks)

Ans: Anscombe's quartet is a collection of four datasets that, when plotted differently on a graph, have the same descriptive statistical characteristics as means, variance, R-squared, correlations, and linear regression lines.

The datasets were developed in 1973 by statistician Francis Anscombe to highlight the value of **data visualization and highlight the limitations of summary statistics** on their own.

Anscombe's quartet serves as an example of the value of exploratory data analysis as well as the limitations of relying just on summary statistics. It also highlights how important it is to use data visualization to identify patterns, outliers, and other important information that summary statistics alone might not be able to provide.

3.What is Pearson's R? (3 marks)

Ans: The test statistic that quantifies the statistical relationship, or association, between two continuous variables is called Pearson's correlation coefficient. Because it is based on the concept of covariance, it is regarded as the best way to measure the relationship between variables of interest. It provides details on the direction of the link as well as the strength of the association, or correlation.

The value of Pearson's R can be between 1 to -1. Where 1 reflects a strong positive relationship and -1 represents strong negative relationship.

3. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans: A technique for normalizing the range of independent variables or data characteristics is feature scaling. It is sometimes referred to as data normalization in data processing and is often carried out at the data preprocessing stage.

Scaling is performed mainly for two reason.
1. The range of raw data varies widely. The target variable will not come properly while running the machine learning algorithm.
2. Gradient decent converges much faster with scaling than without scaling.

Normalization rescales the values into a range of [0,1]. also called min-max scaled. Standardization rescales data to have a mean (μ) of 0 and standard deviation (σ) of 1.So it gives a normal graph.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

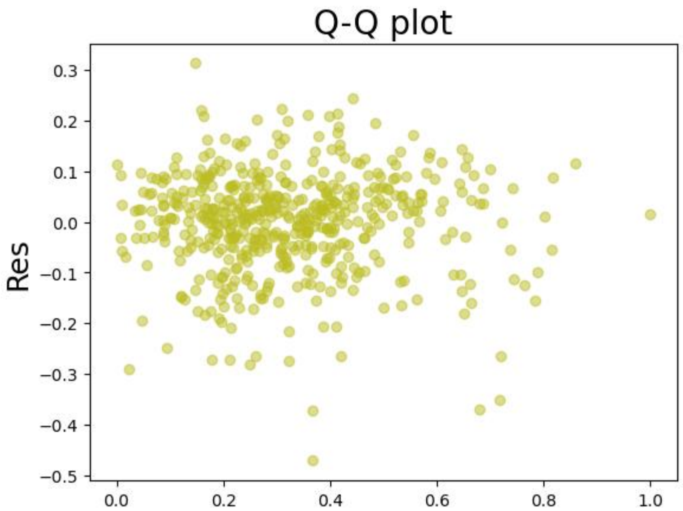
Ans: From the Formula the VIF = (1/(1-R^2)).

So if the R^2 value becomes 1 (one) then the VIF score will be infinite.
R^2 is a measure of the goodness of fit of a model. In regression, the R^2 coefficient of determination is a statistical measure of how well the regression predictions approximate the real data points. An R^2 of 1 indicates that the regression predictions perfectly fit the data.

So in the context of VIF it mean one independent variable perfectly describes by one or more independent variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans: A graphical tool that may be used to determine whether a set of data is likely to have come from a normal or exponential theoretical distribution is the quantile-quantile plot, or QQ plot. A normal QQ plot, for instance, **can be used to verify a statistical analysis's assumption that our residuals are normally distributed**. It is rather subjective because it is only a visual inspection and not a foolproof solution. However, it enables us to quickly determine whether our assumption is reasonable and, if not, how it is broken and which data points are involved.


Q-Q plot
Res
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
0.0
0.2
0.4
0.6
0.8
1.0