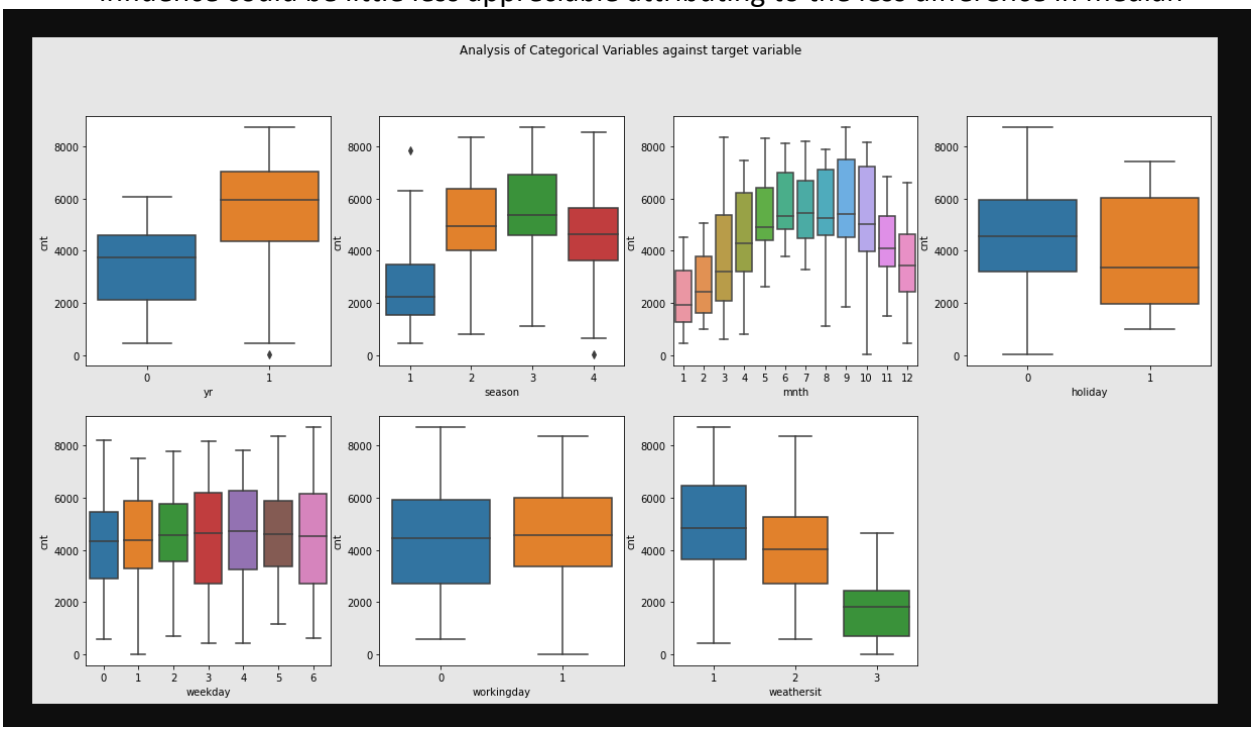
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

From EDA, we could observe as follows:

- ⇒ year seems to have very good influence of total rental bikes as the entire distribution for 2019 is higher than that of 2018 indicating a pattern
- ⇒ season seems to have influence on number of people opting for total rental bikes thereby months also have influence. People tend to opt for more rental bikes in Summer and Fall
- ⇒ workingday & weekday doesn't seem to influence total rental bikes as the median and distribution is similar.
- ⇒ Weather situation seems to influence on total rental bikes where people opt for more rental bikes in clear or partly cloudy weather
- ⇒ People tend to opt for little more rental bikes during no holidays than on holidays. The influence could be little less appreciable attributing to the less difference in median



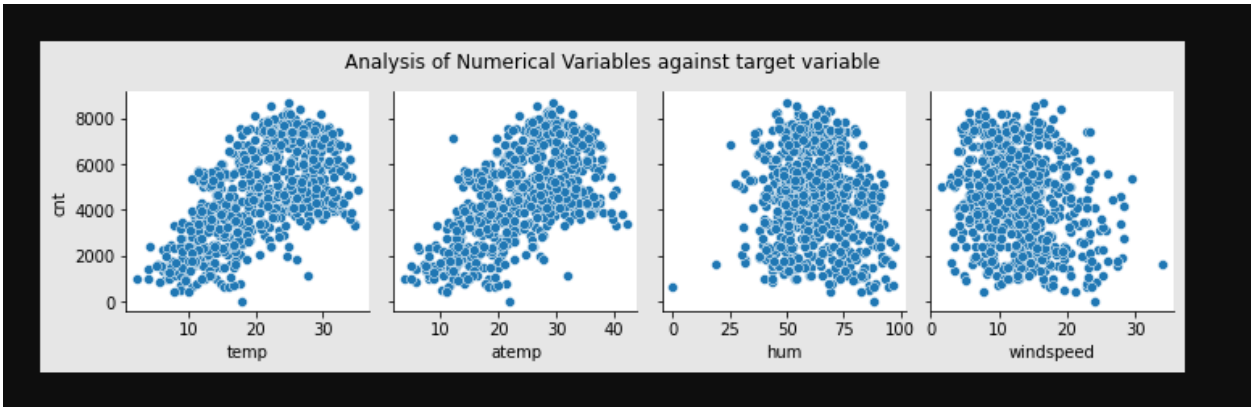
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

- ⇒ It helps to reduce one redundant independent variable so that it reduces the complexity of the model, especially when there are multiple categorical variables with multiple levels.

⇒ It also helps to avoid the correlation among the dummy variables as the removed dummy column is already represented by other dummy columns.

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

Temperature seems to have highest correlation with the target variable. Even though Feeling temperature also have same correlation with the target variable, it is explainable because both temperature and feeling temperature are correlated among themselves



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

⇒ Calculate the Residuals from training target variable and predicted training targer variable and plot the histogram of the same. It should have mean around 0 and should follow normal distribution (approx).

⇒ Plot a scatterplot between residuals and some of the independent variables to see if they have a pattern and also check for change in variance

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

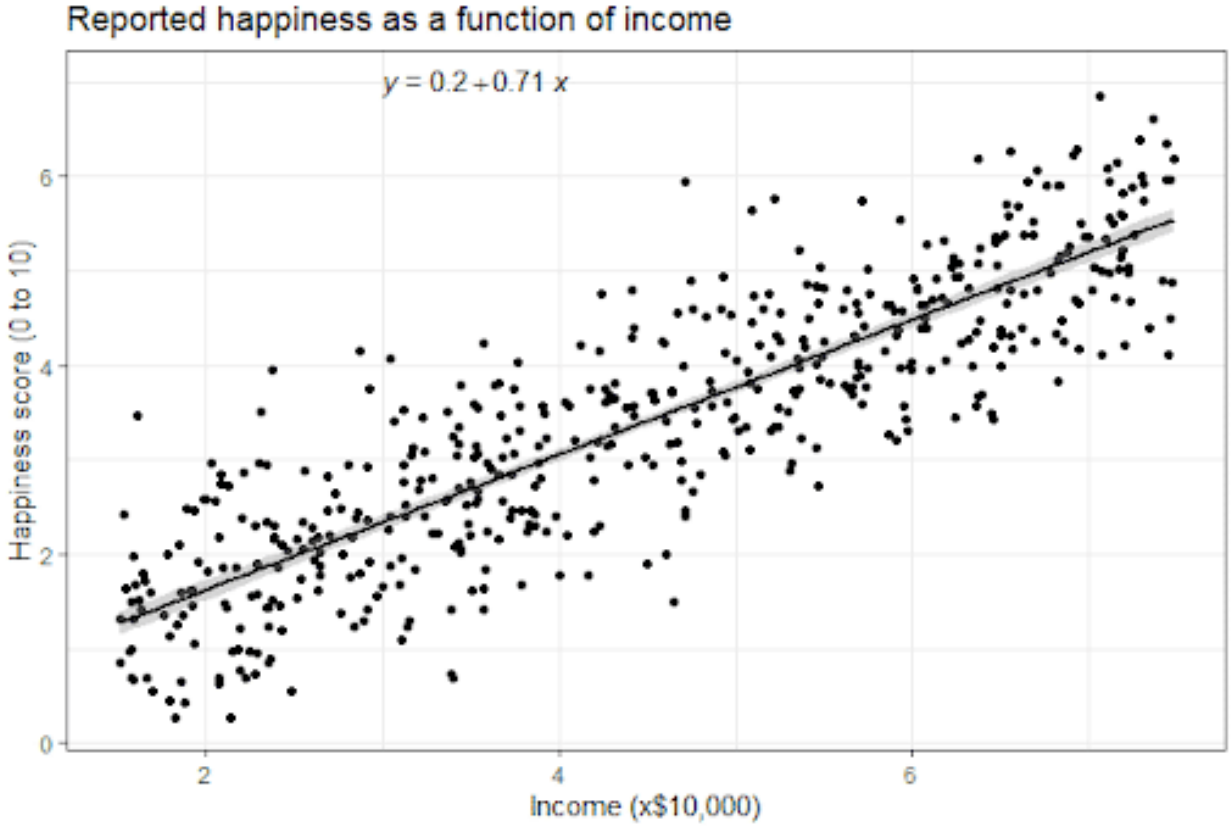
- temperature (Target variable increases by ~0.44 for unit rise in temperature while other dependent variables are held constant)
- Weather Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds Target variable decreases by ~0.30 when weather is light snow while other dependent variables are held constant)
- Year 2019 (Target variable increases by ~0.23 when the year is 2019 while other dependent variables are held constant)

General Subjective Questions

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

Linear Regression is one of the Machine Learning Algorithm which falls under Supervised Learning where the data used for training has defined labels. Regression as such is trying to model a target value based on independent predictor variables. This method is predominantly used for analysing cause and effect relationship between variables. Linear

Regression in particular is the way where we try to model a linear relationship between target values and independent predictor variables.
Simple linear regression is where we have one target variable which is dependent on one independent variables. We try to plot target variable with respect to the independent variable. In almost all cases, we won't get a perfect straight line, target is usually scattered against the independent variable like below



Linear Regression tries to fit a best possible straight line between the target and independent variable which can explain the variance of the data better. The line can be modelled as a linear equation
y = b0 + b1.X
The idea of best fit line is to find the best possible values of b0 and b1 so that the variance is explained. The best fit is accomplished by minimizing the Residual Sum of Squares

Error/Residue for an x ➔ e = y – y_pred (y_pred is the predicted value of target using the best fit line)

Residual Sum of Squares = $e_1^2 + e_2^2 + \ldots + e_n^2$

The strength of the linear regression Model is measured using R^2 which is called coefficient of determination
$R^2 = 1 – RSS / TSS$

Where RSS is Residual Sum of Squares and TSS is Total Sum of Squares which is nothing but the sum of errors of data points from mean.
In simple words, R square = Explained Variance / Total Variance in the data

Multiple Linear Regression is when the model is built using multiple independent variables. It is more often the case that multiple variables are need to explain good variance in the data and better predictions. The model is a hyperplane explained by below equation
$y = b0 + b1X1 + b2X2 + …. + bpXp$
The best fit plane is again accomplished by minimizing RSS.

Below are the assumptions of Linear regression:

1. There should be linear relationship between target variable and some of the independent variables
2. Error terms are normally distributed
3. Error terms have zero mean
4. Error terms have constant variance
5. Error terms should be independent of each other

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

Anscombe's quartet emphasizes the importance of why summary statistics don't tell the whole story about the data and the fact that it can really fool us. Anscombe's quartet comprises of four datasets which are statistically similar. It means it has same variance, mean and so on. But the dataset is peculiar that fools the regression model if built.

Anscombe's Datasets (from Wikipedia):

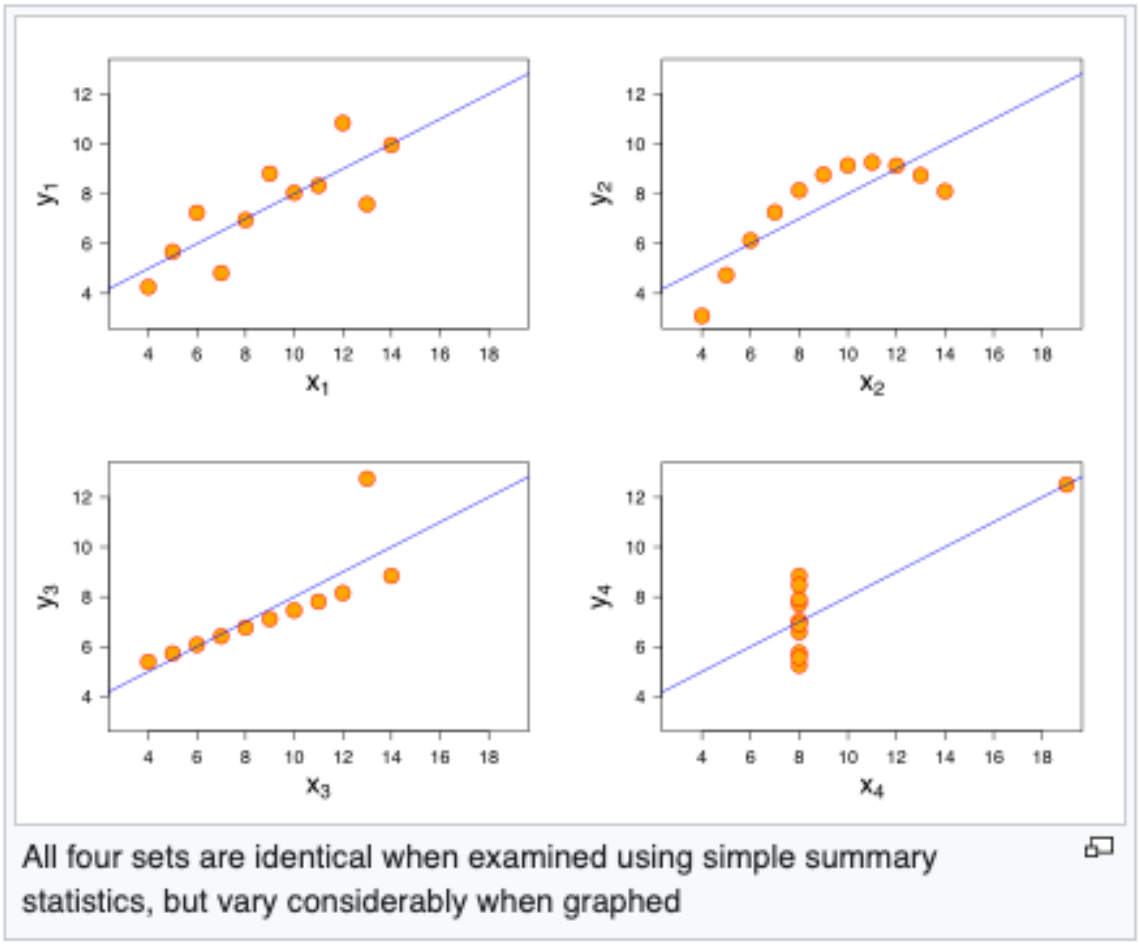
Anscombe's quartet

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

Statistics of four datasets (from Wikipedia):

Property	Value
Mean of x	9
Sample variance of x : s_x^2	11
Mean of y	7.50
Sample variance of y : s_y^2	4.125
Correlation between x and y	0.816
Linear regression line	$y = 3.00 + 0.500x$
Coefficient of determination of the linear regression : R^2	0.67

But when we try to fit a model and plot the a distribution, it can tell a whole different story like below (image from Wikipedia):



If we analyse from graph,

- dataset1 fits the regression model very well.
- Dataset2 is non linear
- Dataset3 has few outliers which doesn't fit the model
- Dataset4 has lot of outliers, model just fits two points

Anscombe's quartet emphasizes the importance of graphing data and tells that basic statistical properties may be inadequate in describing the datasets

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

Pearson's R also called Pearson's Correlation Coefficient is the measure of linear correlation between two sets of data. It is the ratio of covariance of two variables and product of their standard deviations. The value range of Pearson's R is between -1 and 1. For example, if the correlation value of two variables is 0.8 , it signifies that if variable A goes up, variable B will also go up and vice versa. On the contrary, if correlation is negative, it signifies that variable B will go down if variable A increases. As seen by the example, it measures the strength of association and direction between two variables. Please note that Pearson's R doesn't capture any non linear relationship between variables.

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

Scaling is a preprocessing step in model building to change the data to fit a particular range. Scaling is performed mainly for two reasons:

1. Coefficient interpretation becomes easier and comparable across different independent variables
2. The algorithm on the cost function optimization (e.g: Gradient descent) will be optimized if the variables are comparable in magnitude

Normalized Scaling or Min Max scaling fits the range of data between 0 and 1 and it is achieved using the formula

x = x – min(x) / max(x) – min(x)

Standardized Scaling replaces the values with their Z scores. It fits the data such that the mean is 0 and standard deviation is 1. It is given by the formula

x = x – mean(x) / sd(x)

since Normalization fits everything between 0 and 1 , it loses some information like outliers whereas it is not the case with standardization.

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

VIF = infinity = $1 / 1 – R^2$. It means denominator is 0 ➔ $1 – R^2 = 0$ ➔ $R^2 = 1$. It is perfect correlation.

It means the independent variable is accurately explained by other independent variables. If this variable is accurately described by linear combination of other variables, this is redundant and it can be dropped in the model building process.

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

Q-Q plot (Quantile-Quantile plot) are plots of two quantiles against each other. Quantile is nothing but a fraction which divides numerically ordered data into equally proportioned buckets. It is used to find out whether the two sets of data come from same distribution. 45 degree angle is plotted on the QQ plot, if two quantiles fall on that reference line, it means two sets of data came from same distribution.

Since it can confirm whether two sets of data came from same distribution, it has crucial importance in linear regression scenarios where we have test data received separately often after building the model. We can use Q-Q plot to confirm whether the test data and training data are from populations of same distributions. It can also be used to validate:

- if both the populations have common location and scale
- if both the populations have similar tail behavior
- whether the residuals follow normal distribution (Assumption of Linear Regression)