Assignment-based Subjective Questions

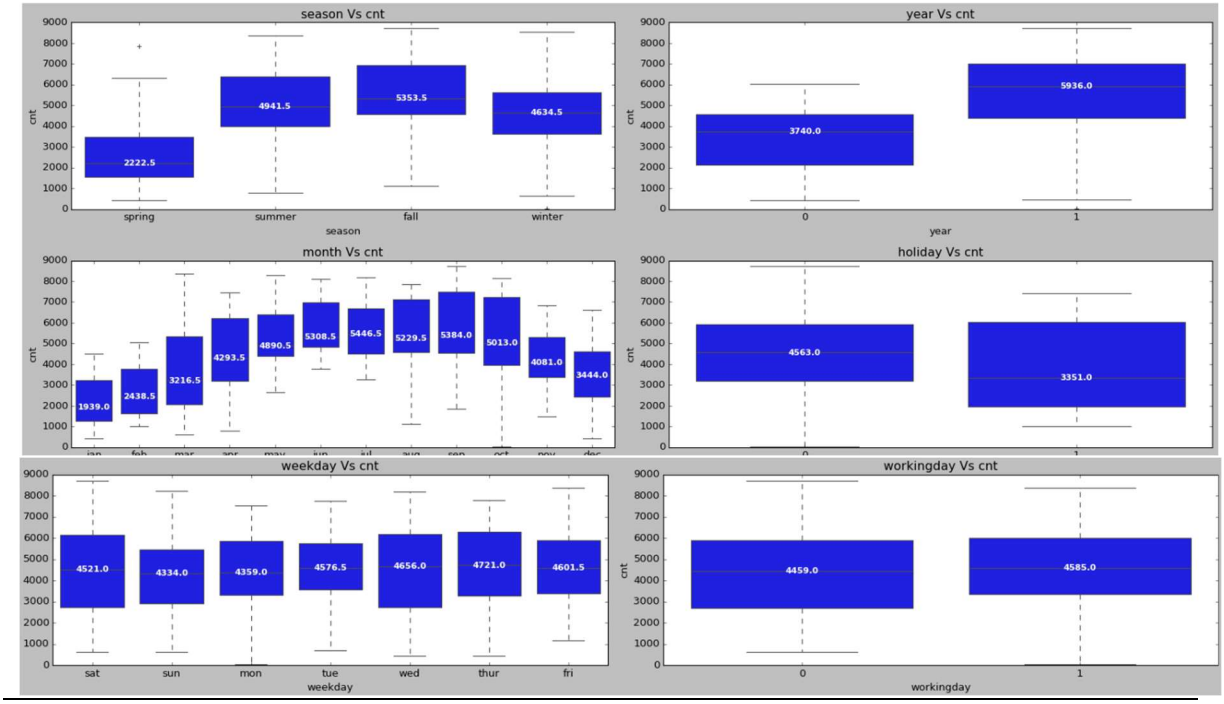
Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

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

There are categorical variables season, year, month, holiday, weekday, workingday which impact the target/dependent variable cnt, as depicted in the figure below, some point to highlight:

- Fall season has the highest bike rentals followed by, summer, winter and spring, since organization should think about expanding during fall, summer and winter, since spring has less booking, this time can be utilized for repairing, services and upgrading the bikes and related infrastructure. Since spring season has a low footfall, organization must think of some offers to attract more buyers during this season.
- June, august, September, October has more demand as compared around other months of the year, whereas January, February, December being the low demand months
- Thursdays and weekends have more demand.
- 2019 saw a sudden increase in the demand of bike rentals as compared to 2018.
- Working days and holidays doesn't have much impact on the bike rentals demands.



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

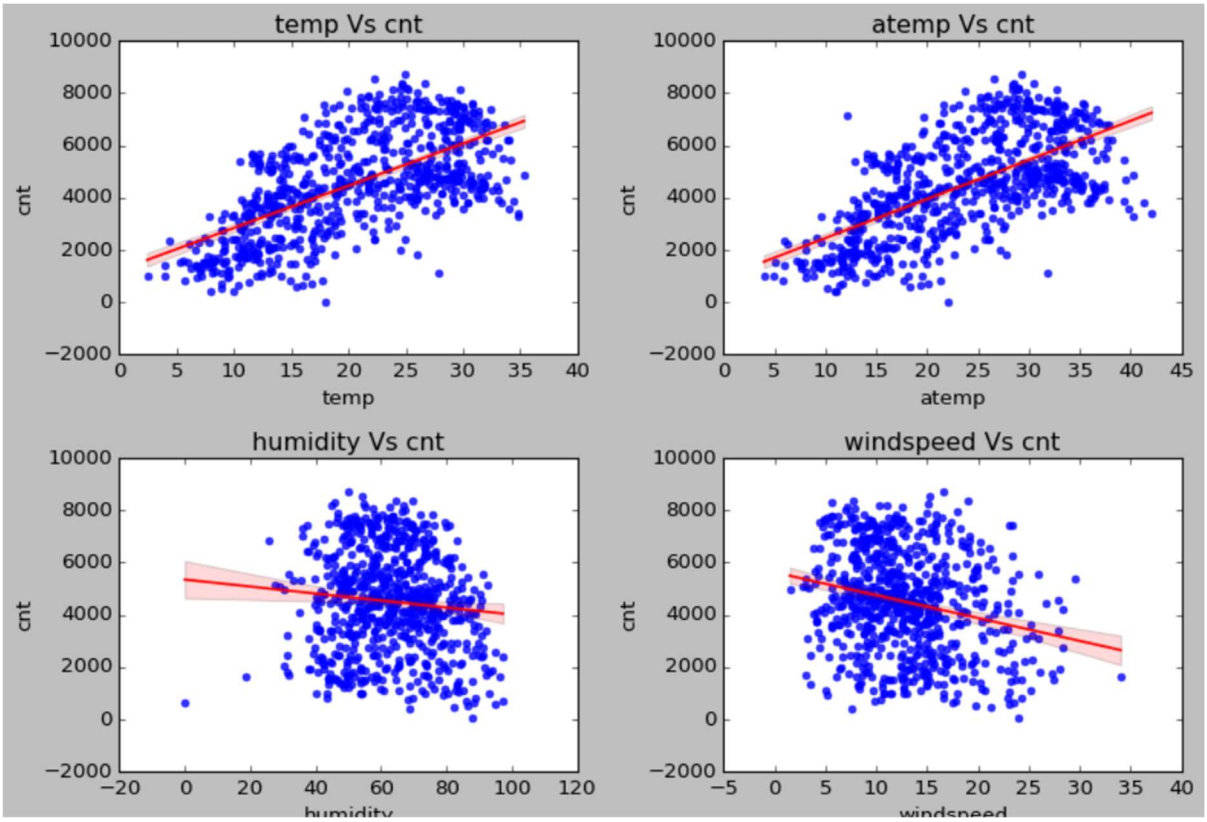
Total Marks: 2 marks (Do not edit)

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

While getting dummies it is always of utmost importance to create dummies to n-1 level given the feature has n levels, as the last or first or dropped column can always be deduced logically and not dropping first would result in redundancy and collinearity, hence drop_first=True is used always.

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)

"temp" and "atemp" variables have the highest correlation to the target variable "cnt", these independent variables have positive correlation with target variable, as compared to other independent variables. Below is the snapshot from the jupyter notebook.



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)

Multiple Linear Regression models are validated based upon fulfilling the criterion specified/based on Linearity of relationships, Independence of errors, Normality of error terms, homoscedasticity (constant variance of error) and No multicollinearity (variance inflation factor).

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)

Top three features are:

1. “temp” or “atemp” (one of both)
2. “season”
3. “year”

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)

<Your answer for Question 6 goes here>

Linear regression is used for predictive modeling used to deduce the linear relationship between a set of independent variables and one dependent variable aka target variable. The relationship is defined by the equation y=mx+c:

Y is the target variable, x is the independent variable and m is the slope/coefficient and the c is y intercept, this equation can be used to describe simple linear regression, multi linear regression is when there are multiple independent variables. c is also called as constant.

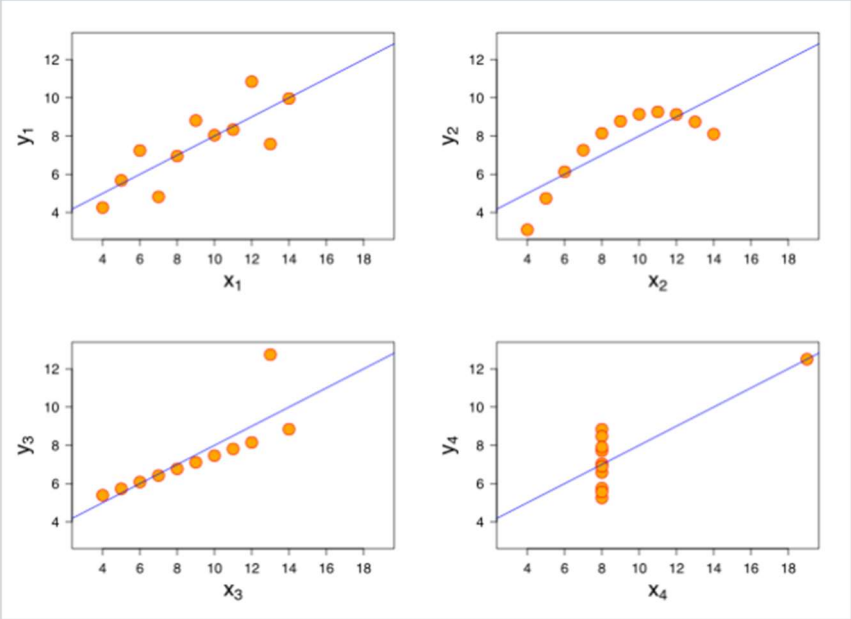
Y= m1.x1+m2.x2+…..+mn.xn+c

These expressions provide the values for the coefficients which describes the relationship between the independent and the dependent variables using the cost function and RSS, TSS, R^2 and adjusted R^2, either positive or negative as the slope of the line in 2D space of x and y coordinates.

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)

<Your answer for Question 7 goes here>

Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different. It is also used sometimes to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties.



The four datasets composing Anscombe's quartet. All four sets have identical statistical parameters, but the graphs show them to be considerably different

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)

<Your answer for Question 8 goes here>

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. When one variable changes, the other variable changes in the same direction.

Below is the formula for calculating the pearson's R:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

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)

<Your answer for Question 9 goes here>

Scaling in machine learning, also known as feature scaling or data normalization, it is the process of transforming a dataset's numerical features to a common scale or range. This is done to ensure that all features contribute equally to the model, which can improve the algorithm's effectiveness and speed up processing.

In both Normalized and the standard scaling, one is transforming the values of numeric variables

so that the transformed data points have specific helpful properties. The main difference is that, in scaling, you're changing the range of your data, while in normalization, you're changing the shape of the distribution of your data.

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)

<Your answer for Question 10 goes here>

This happens during the get dummies where some redundant or collinear data is written as dummy columns e.g. is drop_first is not used, all the features with multiple categories must be converted one by one and then concatenated to form one data frame otherwise it would result in inf VIF.
Secondly if there are columns/features used which are highly correlated and high collinearity is introduced.

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)

<Your answer for Question 11 goes here>

A Q-Q plot, or quantile-quantile plot, is a graphical tool that compares two sets of data to determine if they come from the same distribution. It's a scatterplot that plots the quantiles of one data set against the quantiles of the other.