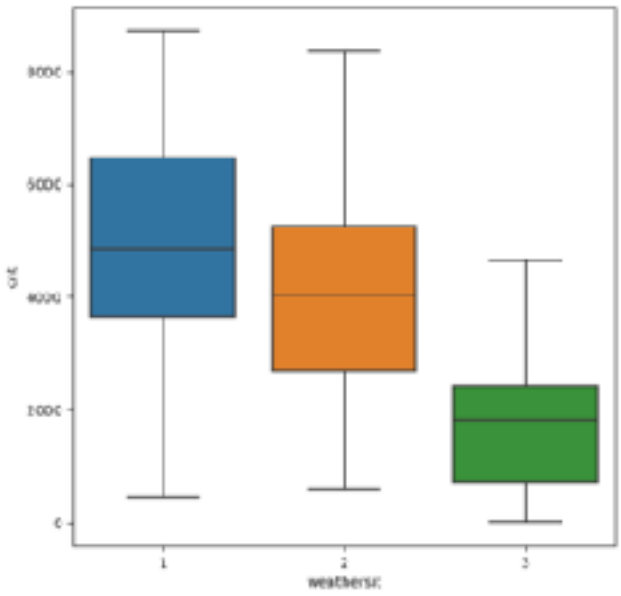
# Assignment-based Subjective Questions

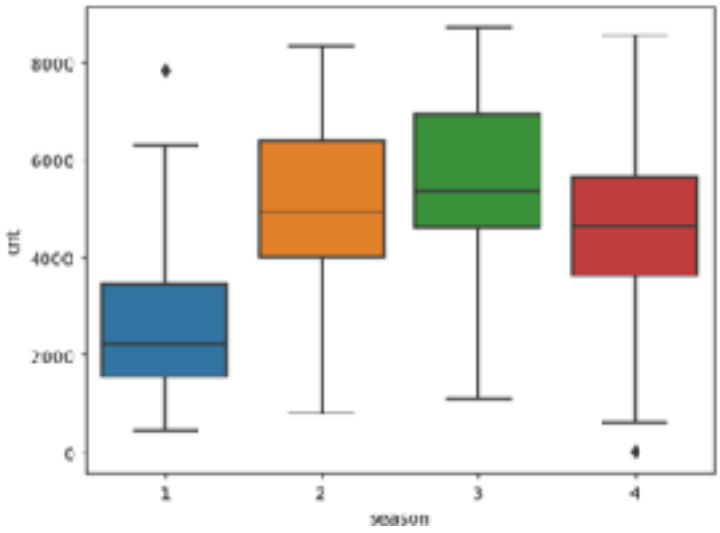
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

   Ans. Season and weather show significant effect on bike demand.



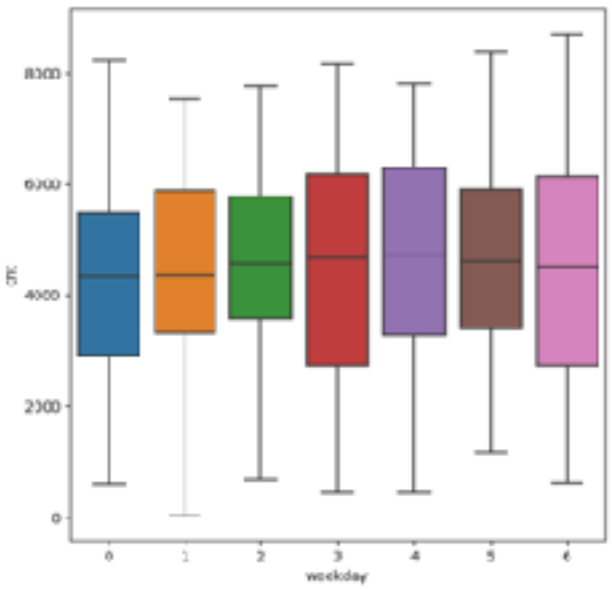
Weathersit - categorical variable on bike demand

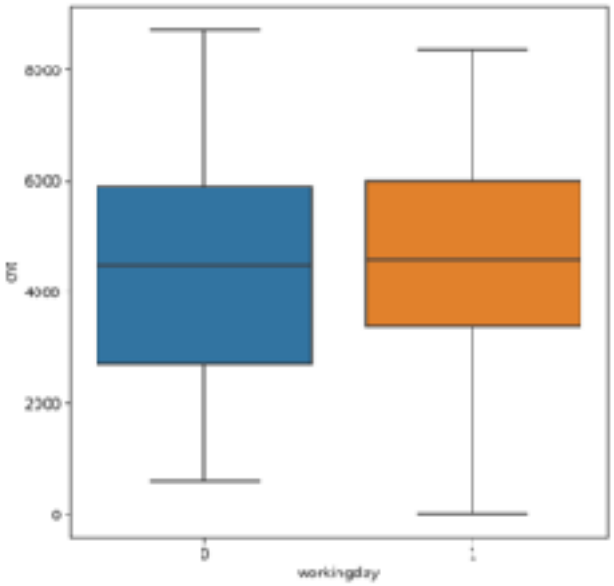
A clear weather shows more demand for bikes, which decreases with increasing clouds and further dips with rain and snow.

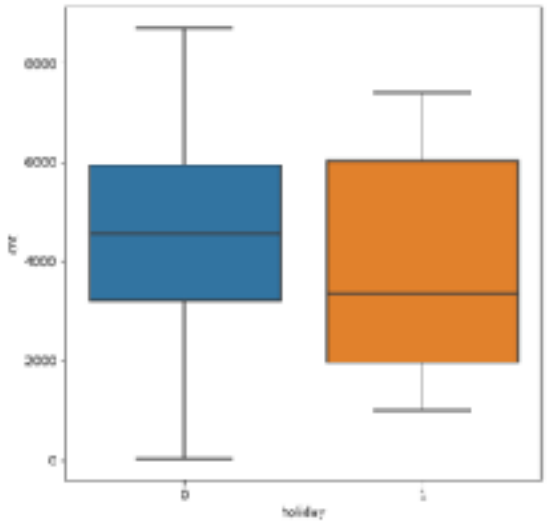



Season on Bike demand

Summer and fall show more demand for bikes compared to  spring and winter season.




While weekday, holiday and working day do not show significant trends on bike demand.

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

Ans. While creating dummy variables, a categorical variable with n levels can be represented by n-1 variables (n-1 variables a sufficient to represent n levels).

But pd.get_dummies by default produces 'n' dummy variables, this would later need us to drop a dummy variable column (to get n-1 variables).

But with drop_first=True the pd.get_dummies function would by default drop the first dummy variable produced. This would result in efficient representation and computation from memory and cpu resource perspective too.

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

Ans. Temp (temperature in Celcius) and atemp (feeling temperature in Celcius) have highest correlation with target variable.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

Ans. One of the assumption is to have a linear relationship between the target variable the predictor this is ascertained through pair plots and correlation matrix, which indicate linear relation between target variable and predictor in this case (For ex: Temp(X) and Cnt(Y))

The second assumption with residual analysis, we define residual/error terms as difference between predicted values by model and target variable. We check for the residual terms to be normally distributed this can be done using a dist-plot.

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

Ans. Temp (temperature in Celcius), windspeed and season (broken down into dummy variable spring, summer, fall and winter) contribute significantly towards demand of bikes.

# General Subjective Questions

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

Linear regression algorithm is a supervised Machine Learning Algorithm, that from a set of labeled input points learns a linear mapping to output/predicted variable(target variable).

In the simplest case a Simple linear regression model tries to obtain a straight line mapping according the data points according equation $y= \beta0 + \beta1 X$. Here the algorithm tries to learn $\beta0$ and $\beta1$ parameters which define the straight line/ mapping.

In complex cases with multiple input predictors, the models tries to learn a hyperplane on which all the input datapoints line and follows the equation.

$$y= \beta0 + \beta1 X1 + \ldots + \beta p Xp + \varepsilon$$

The hyperplane is defined by the parameters/coefficients $\beta0..\beta p$.

The algorithms tries to learn best fit line or hyperplane by minimising the Residual Sum of Squares (RSS), which is the sum of squares difference between the actual target variable in the training dataset and the predicted target variable by the algorithm.

The linear regression algorithm make the following assumptions.

- there exists a linear relationship between the predictor and target variable (X &Y)
- The error in prediction introduced by the model is normally distributed.
- The error terms are independent of each other
- The error terms are have constant variance (homoscedasticity)

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

Anscome Quartet constructed in 1973 by Francis Anscombe is a classic example of how looking at raw data can fool us.

Anscombe's quatret emphasises the need of EDA and data visualization by presenting 4 different datasets having very similar descriptive statistics but when graphed have very different distribution.

The quartet represent 4 datasets each with 11 x-y pairs of data, have similar summary statistics in terms of mean, variance, correlation between x&y and a linear regression line but when plotted each dataset shows a unique connection between x and y with unique variability pattern and distinctive correlation strength.

The quartet lays down the importance of not just relying on plain descriptive statistics, and emphasises the importance of using EDA and data visualization to spot trends, outliers & crucial details that are not obvious from the descriptive statistics.

3. **What is Pearson's R?**

The Pearson's Correlation Coefficient (R) is a common method to measure the strength of correlation between two variable.

The coefficient describes the correlation with a numeric value ranged between [-1,1] indicating both the strength as well as direction of correlation.

A value of 0 indicates no correlation between the variables.

A value between [0,1] indicates a positive correlation in which change in one variable results in change in other variable in same direction (proportional effect- Eg increase in a variable results in increase in correlated variable). A higher value towards 1 indicates strong correlation.

A value between [-1,0] indicates a negative correlation in which change in one variable results in change in other variable in other direction.(non proportional effect- Eg increase in a variable results in decrease in correlated variable). A lower value towards -1 indicates strong negative correlation.

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

Scaling is a data preprocessing step to transform the value of features or variables in a dataset to similar scales/ranges, with an aim to ensure all features contribute equally in a model.

Feature scaling is needed with a model with a number of independent variables on different scales. The resulting model parameters/coefficients would be weirdly varying in different scales leading to difficult interpretation and analysis of the model devised. In addition scaling would help in faster convergence and better model performance when using gradient descent algorithm.

The formula for normalization and standardisation are below

Min-Max Normalisation : x_norm = (x - min(x))/(max(x)-min(x))

Mean Normalisation : x_norm = (x-mean(x))/(max(x)-min(x))

Standardisation : X_standard = (x-mean(x))/standard_deviation(x)

The difference between normalisation and standardisation are:-

- Standardisation is to be used when we want scaled features to have zero mean and unit standard deviation. While Normalisation would yield scaled features in the Range [0,1] or [-1,1] and are useful when features are on different scales.

Standardisation would be useful when feature distribution is Normal or Gaussian, while normalisation is useful when we don't know about distribution.

Normalisation is better in the absence of outliers, while standardisation is less affected by outliers.

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

The VIF is used to detect/explain multicollinearity. It typically involves an estimator model built to predict a independent variable from the set of other independent variables. The process is repeated for each independent variable against combination of others.

VIF is given by the formula.

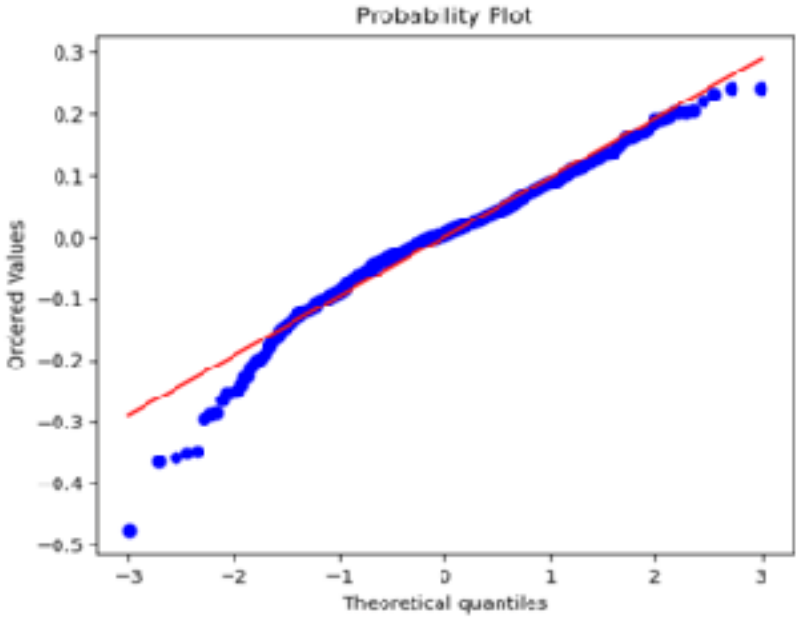
VIFi = 1/1-pow(Ri,2)

VIF will go negative when R_square for a model with the independent variable ‘i’ would be 1 indicating perfect correlation between the variables. In other words the variance in independent variable ‘i‘ is perfectly explained by other variables and the estimator used for computing VIF produces a model with r_square=1(i.e some independent variables create perfect multiple regressions on other independent variables).

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

A Quantile-Quantile (Q-Q) is a scatter plot of quantiles of first dataset against quantiles of other(second) dataset. It is commonly used as a graphical tool to assess if two datasets come from a common distribution, or comparing whether a dataset comes from a theoretical distribution for example like a normal distribution.

In case of linear regression a Q-Q plot can be used in residual analysis to have a graphical view of wether the residual terms conform to a normal distribution. Also in the event of linear regression if train and test data are received separately then Q-Q plot can be used to ascertain both of them are from same distribution.




For example a Q-Q plot from the residual analysis to check if the residual terms are normally distributed. More the points of quantile lying on straight line, more closely the distribution relates to a theoretical normal distribution in the instance of residual analysis.