# Shared Bicycle Demand Prediction Assignment

Mayuresh Kulkarni

mayureshk.123@gmail.com

## 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?
   Relationship of individual categorical parameters with demand can be described as below:
   a. Not many people use bicycle sharing app during spring and winter. They prefer using it in summer and fall.
   b. Many people prefer to use bicycle sharing during non holidays.
   c. Many  Casual user prefer to ride bike on weekends where as registered users prefer weekdays. This makes total bike rides more or less distributed equally throughout the week.
   d. Very few registered users like to ride bicycles on a non working day. For casual users it doesn't matter if it's a working day or not. Therefore total no of bicycles ride taken on working days is more.
   e. Few people ride bicycles in the month of Jan to March and November December. Demand is really high in rest of the months. This variable has been treated as ordinal variable to capture seasonality in demand forecast.
   f. More people prefer using bicycles on a clear or cloudy day rather than a misty day. Very few people ride bicycles on a day with light snow or rain. Whereas there is no one riding a bicycle on a day with heavy snow or rain or thunderstorms or thick fog. Notice that eventhough the weather situation is a categorical variable, it is treated as ordinal variable as the value increases with increasing severity of the weather.

2. Why is it important to use drop_first=True during dummy variable creation?
   It is necessary to use drop_first = True as the categorical variable used in the data have mutually exclusive values. What it means is that it could have only one value. So for instance consider categorical variable season, it could have either of values 1, 2, 3, 4. So the four seasons can be represented as TFFF, FTFF, FFTF, FFFT, where F is false and T is true. Now consider that we drop the season 1 from representation, we can still all the season without the need to use another column for season 1. FFF, TFF, FTF, FFT is now used to represent all the four seasons. This helps in saving a lot of memory and runtime in the analysis.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   Looking at the plots, Temp and feel temperatures have highest correlations to the no of bicycle rides.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   Since data is split between training and testing data, the model is created based on training data. Next it is validated on testing data using R2 Score. Since the R2 score is 0.82 on

testing data set it is considered a good fit.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   Temperature, Feel Temperature, Holiday, Month, Year, Working day, season and weather situation contribute significantly to prediction of demand for bicycle ride sharing app.

## General Subjective Questions

1. Explain the linear regression algorithm in detail.
   Linear regression is a supervised machine learning algorithm used to model the relationship between a dependent variable (target) and one or more independent variables (features) by fitting a linear equation to observed data.

   For a single feature (simple linear regression):

   $y = w_0 + w_1 x$

   y: Predicted value (target)
   x: Feature (input variable)
   w_0: Intercept (bias)
   w_1: Slope (weight)

   For multiple features (multiple linear regression):

   $y = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$

   The goal is to find the weights (( w_0, w_1, ..., w_n )) that minimize the difference between the predicted values and the actual values in the training data.

   This difference is measured using a loss function, typically the Mean Squared Error (MSE):

   $$MSE = \sum \left(y_i - y_i^{'}\right)^2$$

   y_i: Actual value at ith index

   y'_i: Predicted value at ith index

   Initially we start with a random weights usually 0. Then based on the MSE values we adjust the weights using gradient descent method for optimization. And then we repeat the process until we receive minimum MSE value.

2. Explain the Anscombe's quartet in detail.
   Anscombe's quartet is a collection of four different datasets that have nearly identical simple statistical properties (such as mean, variance, correlation, and linear regression

line), but appear very different when graphed. Hence it is necessary to graph out and visualize the data before constructing the model.

3. What is Pearson’s R?
   Pearsons R is a correlation coefficient that measures linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationships or correlations. As a simple example, one would expect the age and height of a sample of children from a school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   Scaling is the process of transforming features in your dataset so that they have similar ranges or distributions. This is important because many machine learning algorithms perform better or converge faster when features are on comparable scales.
   **Why is scaling performed?**
   a. Improves model performance: Algorithms like k-NN, SVM, and gradient descent-based models are sensitive to the scale of input features.
   b. Prevents dominance: Features with larger ranges can dominate distance-based or weight-based calculations.
   c. Speeds up convergence: Helps optimization algorithms converge faster.

   **Types of Scaling**

   **1. Normalized Scaling (Min-Max Scaling)**

   **Definition:** Rescales features to a fixed range, usually [0, 1].

   **Formula:**

   $$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

   **Use case:** When you want all features to have the same scale and you know the minimum and maximum values.

   **2. Standardized Scaling (Z-score Standardization)**

   **Definition:** Transforms features to have a mean of 0 and a standard deviation of 1.

   **Formula:**
   $x' = \frac{x-\mu}{\sigma}$

where $\mu$ is the mean and $\sigma$ is the standard deviation.

**Use case:** When features are normally distributed or when you want to center and scale features.

5. .You might have observed that sometimes the value of VIF is infinite. Why does this happen?
The value of VIF (Variance Inflation Factor) becomes infinite when there is perfect multicollinearity among the features. This means that one feature can be expressed as an exact linear combination of one or more other features in your dataset.
The denominator in the VIF formula is based on ( $1 – R^2$ ) from regressing one feature on all others.
If a feature is perfectly predicted by the others (( $R^2 = 1$ )), then ( $1 - R^2 = 0$ ), and the VIF becomes infinite (division by zero).

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, most commonly the normal distribution.

   Use in Linear Regression
   In linear regression, one key assumption is that the residuals (errors between predicted and actual values) are normally distributed. A Q-Q plot helps you visually check this assumption.

   How to use:
   Plot the quantiles of your residuals against the quantiles of a normal distribution.
   If the points fall approximately along a straight line, the residuals are normally distributed.
   Importance
   Validates model assumptions: Ensures that the normality assumption for residuals holds, which is important for reliable confidence intervals and hypothesis tests.
   Detects deviations: Helps identify skewness, heavy tails, or outliers in the residuals.
   Model diagnostics: If the Q-Q plot shows strong deviations from the line, you may need to transform your data or use a different model.