Assignment based Questions – Answers

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**
   There were six categorical variables. From the box plot we can infer the below:

   a. **season**: Almost 32% of the bike booking were happening in season3 with a median of over 5000 booking (for the period of 2 years). This was followed by season2 & season4 with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable.

   b. **Mnth** : Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.

   c. **weathersit**: Almost 67% of the bike booking were happening during 'weathersit1 with a median of close to 5000 booking (for the period of 2 years). This was followed by weathersit2 with 30% of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.

   d. **holiday**: Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.

   e. **weekday**: weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor.

   f. **workingday**: Almost 69% of the bike booking were happening in 'workingday' with a median of close to 5000 booking (for the period of 2 years). This indicates, workingday can be a good predictor for the dependent variable

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

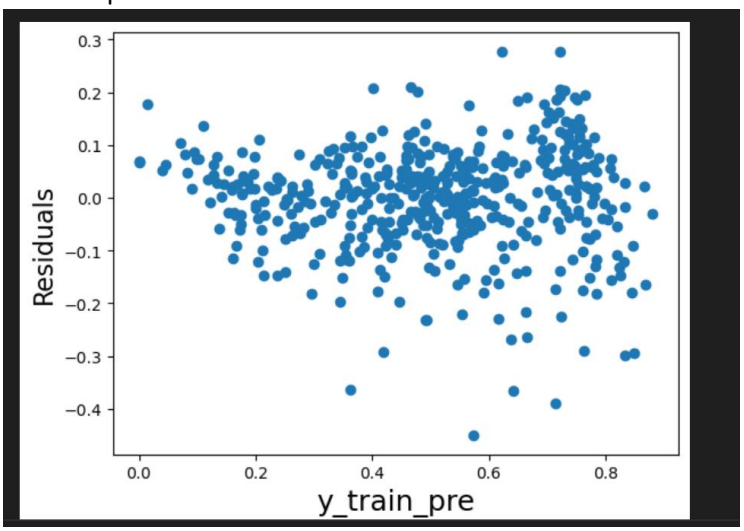
   drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
   Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished. Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   From the pairplot, temp and atemp are highly correlated with the target variable cnt

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

   After reaching the best model, to validate assumptions

   1. Errors are normally distributed – histogram is plotted with the final data set to validate this assumption
   2. Multicollinearity – from the VIF values obtained form the last model, we can clearly see that there is no multicollinearity as the VIF values of all predicting variables are now less than 3
   3. Homoscedasticity – Error terms have constant variance – by a scatter plot of predicted values of y in train data set and the corresponding residues we could visualize how the errors terms are spread. From the figure below, errors terms are spread with constant and independent of each other



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

Final Observations:
temp, yr, weathersit_3 and windspeed are the top four varibles impacting the demand for bikes

When temp increases by 1 unit demand for bikes increases by 0.5661 numbers
When yr increases by 1 unit demand for bikes increases by 0.2339 numbers
When weathersit_3 increases by 1 unit demand for bikes decreases by 0.2491 numbers
When temp increases by 1 unit demand for bikes decreases by 0.1520 numbers

General Questions – Answers

1. **Explain Linear regression algorithm**

   A linear regression algorithm explains how independent and dependent variables are related to each other. Linear regession is used to predict numerical variables. Below are the steps:

   a. Data preparation, cleaning, dealing with categorical varible creating dummies wherever required
   b. Split the data into train and test sets
   c. Split the train set to X(predictor variables) and Y(final numerical variables being predicted)
   d. Fit a linear model using python stats model library. The regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.
   e. In case of one feature the linear line is fitted using below equation
      $Y = \beta0 + \beta1x$
      In case of multiple variables,
      $Y = \beta0 + \beta1x + \beta2x + \beta3x + \ldots. + \beta nx$
   f. Once we get the best values for β0 and β1, we get the best fit line. Using this we can predict the values of y in the test data set

2. **Explain Anscombe's quartet**

   Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

   There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets. This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc

   The four datasets can be described as

   Dataset 1: this fits the linear regression model pretty well.

   Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.

   Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model.

   Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model.

3. **What is pearson's R**

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

"Tends to" means the association holds "on average", not for any arbitrary pair of observations, as the following scatterplot of weight against height for a sample of older women shows. The correlation coefficient is positive and height and weight tend to go up and down together. Yet, it is easy to find pairs of people where the taller individual weighs less, as the points in the two boxes illusstrates.

The Pearson's correlation coefficient varies between -1 and +1 where: r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction) r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

r = 0 means there is no linear association

r > 0 < 5 means there is a weak association

r > 5 < 8 means there is a moderate association

r > 8 means there is a strong association

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

Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm. Scaling is done only for numeric variables

The two most common used scaling methods are:

A. Normalization/Min-Max Scaling: It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python
B. 2- Standardization Scaling: Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

- The variance inflation factor (VIF) quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing collinearity/multicollinearity. Higher values signify that it is difficult to impossible to assess accurately the contribution of predictors to a model. VIF = 1/1-R^2 If there is perfect correlation, then VIF = infinity.

A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. This would mean that that standard error of this coefficient is inflated by a factor of 2 . The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity.

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

Q Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution.

A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line. The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

The slope tells us whether the steps in our data are too big or too small . for example, if we have N observations, then each step traverses 1/(N-1) of the data. So we are seeing how the step sizes (a.k.a. quantiles) compare between our data and the normal distribution.