Linear Regression Assignment

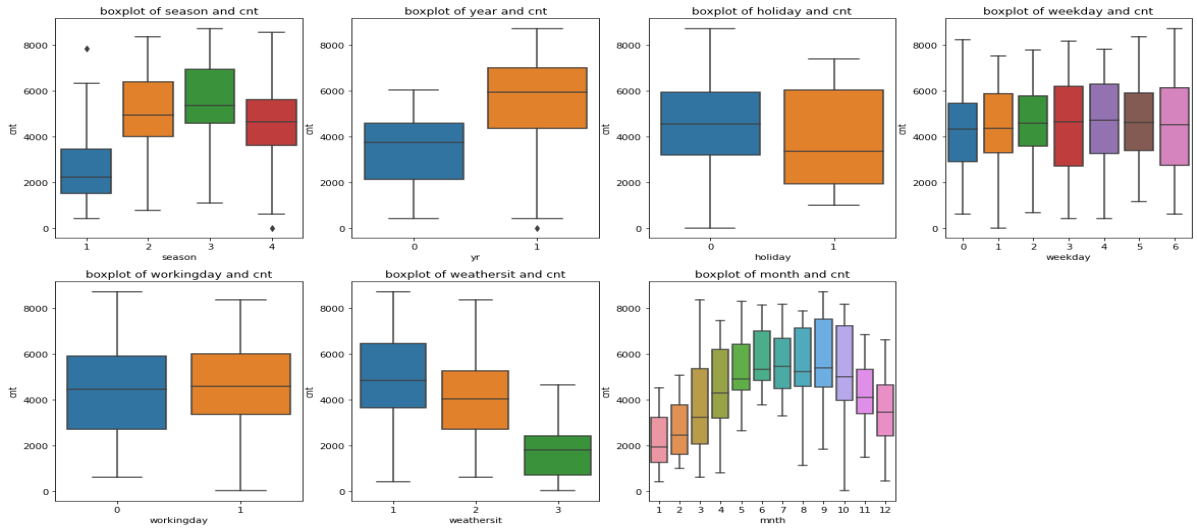
Assignment-based Subjective Questions :

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

Ans) Based on our analysis only some of categorical variables have effect on dependent variable cnt. Based on our final model year, holiday, and some part of season, month and weather situation effect dependent variable. Year, month and season have positive effect while holiday and weather situation has negative effect. Below is the final model equation

cnt = 0.278 + 0.229*yr - 0.092*holiday + 0.632*atemp - 0.290*hum - 0.164*windspeed + 0.118*season4 + 0.074*month5 + 0.091*month9 - 0.184*wtst3

Graphically also we can see:



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

Ans) Generally We want the model to have less variable(lean model) and maximum explanation because it will be easy to interpret. Now if we have a categorical variable with n levels we can explain all n categories using n-1 features. Because one variable can be base variable or base state which has values zero. So we can drop one, that's why we use drop_first as True.

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

Ans) Based on pairplot with respect to target variables temp and atemp have highest correlation with target variable. However temp and atemp both are also highly correlated so you can choose any one.

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

Ans) Broadly there are 4 assumption:

a) Linear relationship with target variables that we checked using scatter plot.
b) Residuals have normal distribution with mean 0 that we check by finding residuals for all points of training set(y_train – y_train_pred) and plot its histogram.
c) Residuals independence, for this we plot residual with y_train_pred and saw it was random.
d) Homoscedasticity, For this we plotter residual with y_train_pred and saw that there is almost constant variance.
e) No multicollinearity and No collinearity, we checked this during model building.

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

Ans) Based on final equation

cnt = 0.278 + 0.229*yr - 0.092*holiday + 0.632*atemp - 0.290*hum - 0.164*windspeed + 0.118*season4 + 0.074*month5 + 0.091*month9 - 0.184*wtst3

The top 3 important features based on coefficients are

a) atemp(feeling temperature in Celsius)
b) Hum(Humidity)
c) Yr(Year 2018 or 2019)

General Subjective Questions

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

Ans) Linear Regression is supervised machine learning algorithm where target variable is continuous(numeric). In Linear regression we try to find the best line possible to predict dependent variable from independent variable.

There are two types of linear regression:

a) Simple linear regression: In this eqn is y=c+mx1(Only one independent variable)
b) Multiple Linear Regression: In this eqn is y=c+m1x1+m2x2+m3x3+…. (Many independent variable)

So in linear regression there is a notion of finding best fitting line which can be found out by finding residual sum of squares which is also cost function of linear regression.

RSS= sum of squares of (y_actual -y_predicted(by model))

So it tries to find coefficient in such a way to minimise this RSS.

2. Explain the Anscombe's quartet in detail.

Ans)Usually People think descriptive statistics tell us everything about data and we don't need visualisation. This is where anscombe's quartet comes into picture. He took 4 different datasets of two variables x and y and shown that all 4 datasets have same descriptive statistics like mean, variance, correlation etc and have same regression line. However if you plot them on graph all 4 shows different behaviour. Which means that sometimes decision based on only descriptive analytics can be misleading.

- Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.
- Data-set II — shows a curve shape but doesn't show a linear relationship (might be quadratic?).
- Data-set III — looks like a tight linear relationship between *x* and *y*, except for one large outlier.
- Data-set IV — looks like the value of *x* remains constant, except for one outlier as well.

3. What is Pearson's R?

Ans) Pearson's R is the pearson correlation coefficient and it 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.

Between 0 and 1 : when one variables changes, other also changes in same direction.

0: No relation

Between -1 and 0: : when one variables changes, other changes in opposite direction.

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

Ans) Scaling is the process of transforming all the numerical variables in same range so that it will be easy to interpret them. Scaling is performed mainly because of two reasons, One is so that the coefficients becomes more interpretable and second is it helps in faster convergence of gradient descent.

Normalised Scaling: It scales the variables in a way that the range is compressed between 0 and 1. x_normalised= (x-xmin)/(xmax-xmin.)

Standardisation: It scales the variables in such a way that mean becomes zero and standard deviation becomes 1. X_standardised = (x-mean(x))/std.dev(x)

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

Ans) VIf=1/(1-rsquare), so if vif is equal to infinity it means rsquare is equal to 1. And if rsquare of a variable is 1 it means that independent variable is explained completely by combination of all other independent variables. It means that variable is multicollinear. So if vif is infinity it means u can remove that feature from your model as it is only increasing noise and not increasing model power**.**

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

Ans) Quantile-Quantile (Q-Q) plot, is a graphical way to checkwhether if a set of data came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

It is used to check whether two data sets:

a) Came from same distribution

b) Have common location and scale

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. And if all the point lies on approximately straight line then it means that both sets have similar distributions.

In linear regression, when we have training and test data set received separately then we can confirm using Q-Q plot that both the data sets are from populations with same distributions or not. This is how they can be used in linear regression.