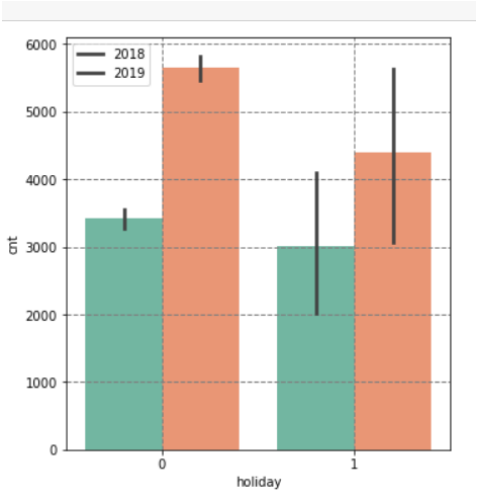
By Sailesh Bathala

Assignment-based Subjective Questions

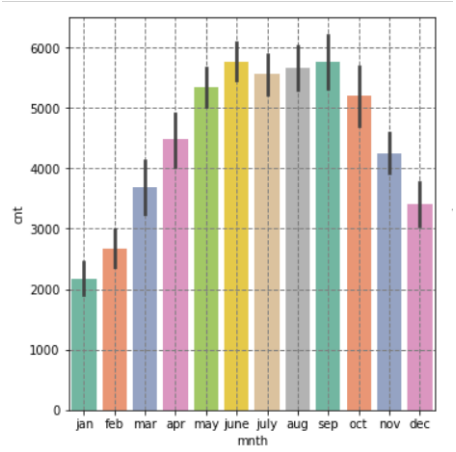
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

   Based on the analysis done by me on the categorical cols I concur to the following:

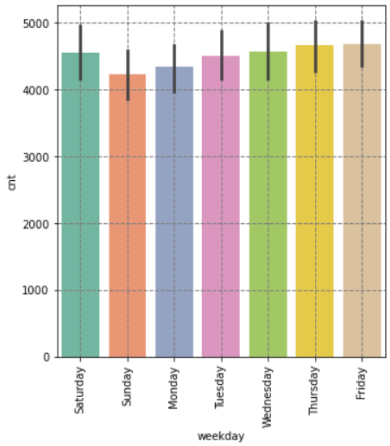
   a. Bookings were less on holidays compared to working days, which could be due to usage of the bike to commute to work and a possibility of staying home or an outing during holidays.



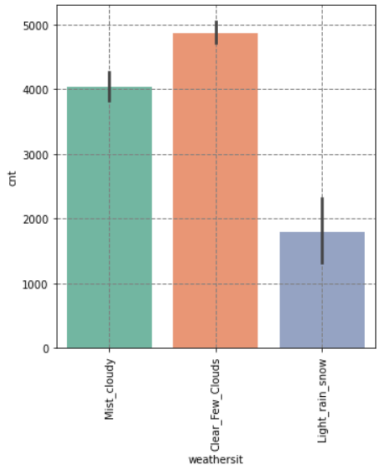
   b. We can concur that, from May to October, the booking trend increased, mostly due to pleasant weather during that time of the year.



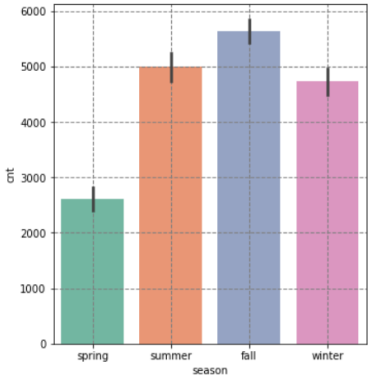
c. We can concur that, towards the end of the week, booking trend increases than compared to the start of the week.



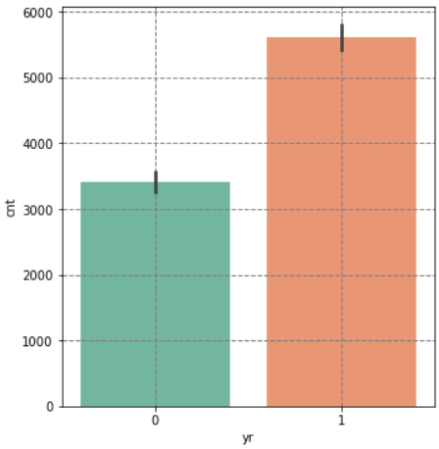
d. We can concur that when the weather is clear, we have more bookings, which is obvious as it is more pleasant to ride a bike during that time.



e. We can observe from the above that, during 'fall' season, we have higher bookings/ridership compared to other seasons.



f. As we have already seen in our previous analysis of other predictor variables, it is seen that 2019 has attracted more bookings than 2018. This could be due to many factors, but it is clear that the company is gaining popularity.



2. **Why is it important to use drop_first=True during dummy variable creation?**
   It is very important that when we create a dummy variable, there is a possibility of high correlation due to this. In order to avoid this, we use drop_first=True to reduce the column created due to dummy variable. The main way to reduce the dummy variable column is to drop one of the categorical columns, i.e., if there are m categories then use m-1 in the model.

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

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

   Did the following to conclude my assumptions:

   a. Plotted a heat map to determine the multicollinearity amongst different variables. They should be less significant.
   b. Linearity should be present among the variables. I could determine this to validate my assumption.
   c. We should make sure that the error terms should be normalized. Only in this case we can conclude our assumptions.
   d. Ensure that we have a VIF of less than 5 and also the p value to be nearing 0.

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

   The top three features that significantly contributed to explaining the demand of the shared bikes is

   - 'year' : This showed significant contribution towards the demand of bikes, where the ridership has seen a huge spike in 2019 compared to 2018.
   - 'Temp' : This showed a lot of significance towards the demand of bikes, where in, more optimum the temperature, more the demand.
   - 'winter' : This played a big role in determining the demand of shared bikes.

General Subjective Questions

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

   Linear regression algorithm is a machine learning algorithm. This is mainly based on supervised learning. It works on the relationship between one or

more predictor variables with one target variable. It is commonly used for predictive analysis of the generated model. If you consider the case study, we were able to quantify the relation or impact of temp, winter, holidays etc, that are the predictor variable, on the cnt(bookings) which is the target variable. There are a set of assumptions that are taken into consideration while performing Linear regression i.e., all the required variables are included and part of the analysis. Addition of other variables could make certain existing variables less significant and also could increase the standard error by a huge margin. One more such assumption is that the data always follows normal distribution.

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

   Anscombe's quartet as the name suggests, has 4 datasets that are nearly identical descriptive statistic but end up giving weird distributions and show up differently on a scatter graph. Anscombe's quartet gives us a picture of how important visualization is before applying them to algorithms and models. So, before we actually interpret the data or probably apply machine learning algorithms to the data, we should first visualize it and then go ahead with the analysis to get a well fit model.

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

   Pearson's R, also known as Pearson's correlation, is commonly used in LR. Correlation coefficients are usually used to measure the relationship between 2 variables. It basically tells us if we can draw a line graph to represent the data. It can range between these three different answers, +1(positive correlation), 0(No correlation) and -1(Negative correlation). One main advantage here is that the two variables being measured need not be of the same unit. One disadvantage is that it cannot determine the slope.

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

   Scaling is a method used to rescale the values to a much easier range say 0 and 1. This makes it easier to compute higher values with more ease and also make the computation faster. Now consider a dataset given. It could have values with high magnitudes with lower units or range. This could give rise to incorrect modelling. So, the best solution here would be to bring it to scale and hence of measurable units. Scaling doesn't affect any coefficients or any other computed values.

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

   This could happen when we have very high correlation which is also termed as perfect correlation. This is nothing but us getting an R2 value of 1 which will give us infinity(1/(1-R2)). As, we have seen in the case study, when we get high VIF, we drop certain selected variables from the dataset which could be causing this. Once we drop such a variable, we will notice that the VIF will slowly start to come down.

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

   Q-Q plots are also known as Quantile-Quantile plots. Here we compare the quantiles of theoretical values with the quantiles of a sample distribution. It helps us determine if the dataset belongs to a normal, exponential or uniform distribution. It can be used with a large population or even sample size data also. It can also determine the change in scale and location.