Assignment-based Subjective Questions

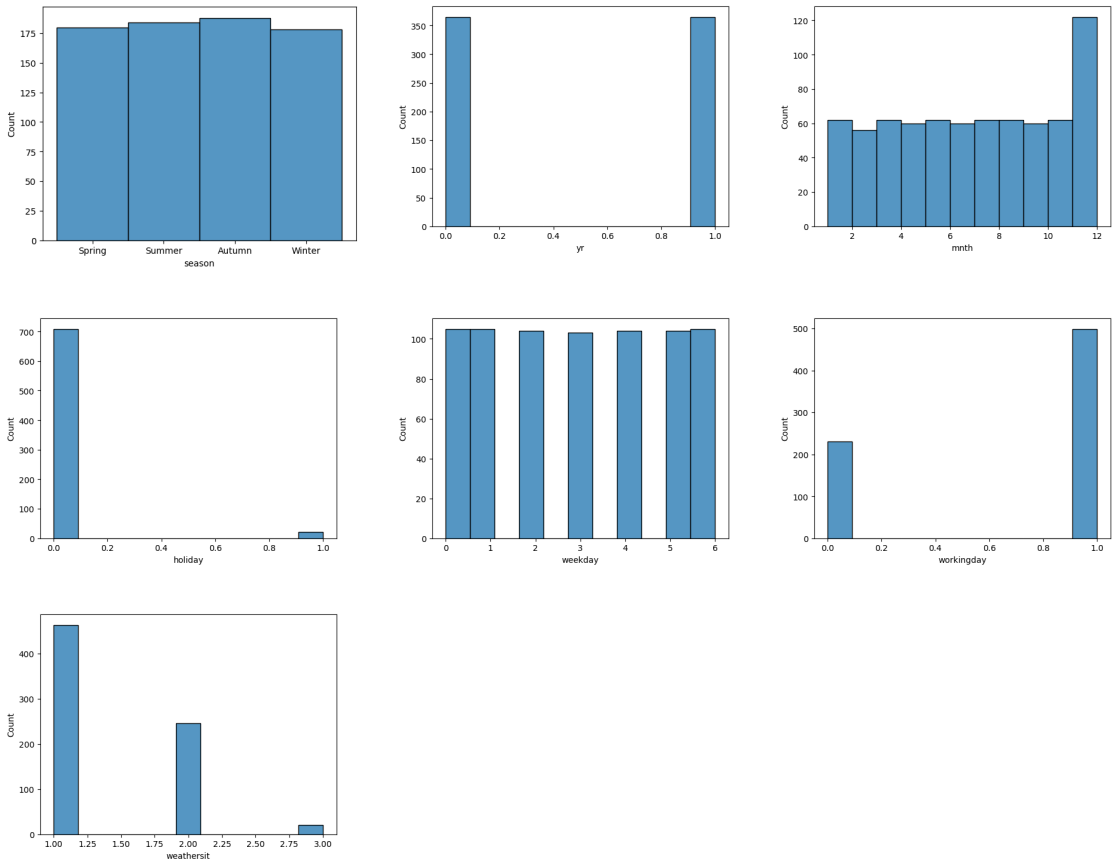
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

   Ans : To understand the characteristics of the dataset. I have performed the EDA, which gives more of a visual understanding of how each variable is related and distributed.

   The categorical variables that i have considered are,

```
cat_col =
['season','yr','mnth','holiday','weekday','workingday','weathersi
t']
num_col = ['temp','atemp','hum','windspeed','cnt']
```

   I have been observing the data set with nunique() and dtypes methods before this categorisation. Let me put some insight that i got from univariate analysis of categorical variables below first,



Looking at this plot, the **weekday** and **yr** has very less effect on the count, the dependent variable. **Holiday, working day** and **weathersit** is likely to have some significant relation to the count as the count varies significantly along with the changes in the x axis. The season vs count also shows some effects on the count value but I doubt whether it is significant or not.

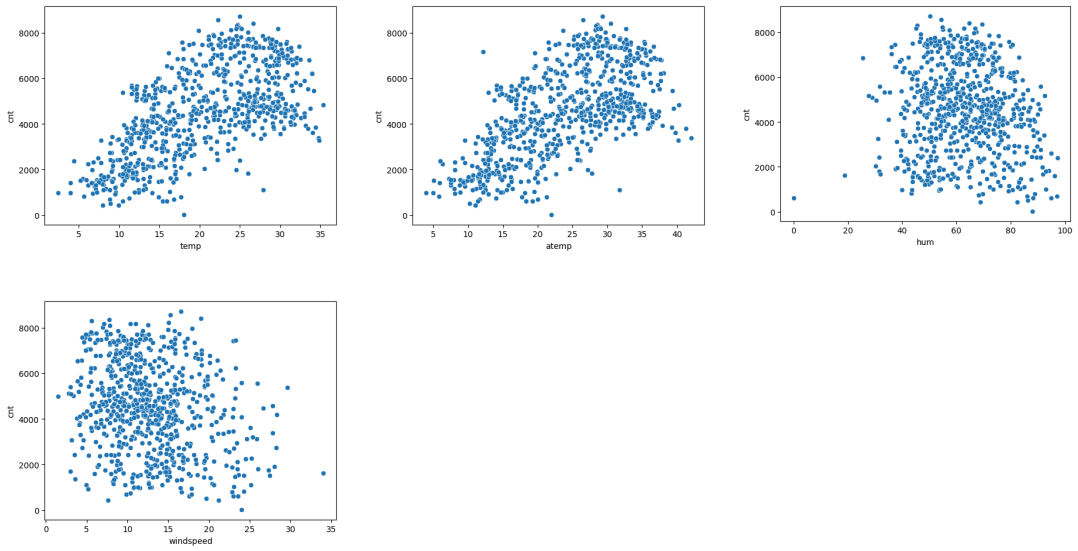
The plot of **mnth vs count** feels like there are some outliers in the data. The last count in the month 12 seems very different from all other data. This can be an outlier or a significant change due to season or some other reason.

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

   Ans : It is important to use drop_first = True to avoid multicollinearity.
   When we hot encode for 'n' categories in a feature, we essentially create 'n' different columns in the dataframe. And these columns will have a relation to each other.
   Let us take our case of seasons, we are hot encoded with columns , Spring, Winter, Autumn and Summer. If the value is spring on one data point all the other will be 0. Dropping one column in this context will eliminate the possibility of predicting the other values in the columns, here the Autumn is eliminated

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

   Ans: Here is my bivariate analysis plots,



Here the temp and atemp seems to have positive correlation with the depended variable. There is an increasing trend as the x value goes from right to left. The hum and windspeed seems to have very less correlation, the data points are scattered and doesn't have a pattern like the other two.

Both temp and atemp seems to have high correlation with the target variable. Looking at them more closely, atemp has some more scattered points other than temp, and may be this might be an indication of a highly correlated feature .

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

   Ans: The one validation that i can state here is the VIF calculation that we did. We have calculated the variance inflation factor for independent variables and found that the values are in the range of 1 - 1.6 which is way lower than 5. Which proves that there is no multicollinearity among them.

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

   Ans: The Magnitude of coefficients generally imply larger impact on the target and P values indicate the statistical significance. So looking at this two values of each features the top 3 continuous features are,
   1. 'Yr' has highest positive coefficient and very low P value
   2. 'Temp' has second highest positive coefficient value and low p value
   3. 'Weathersit' has the highest negative coefficient and low p value among continuous variables

   Spring has the highest coefficient value , but since it is a hot encoded value, i don't know whether it can be considered in the top 3 list.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Ans: Regression is a supervised learning technique which can predict continuous variables based on independent input variables. Linear regression computes the linear relation between the dependent/targeted variable and one or more independent variables / features.
   Simple Linear Regression:
   When there is only one independent variable, it is a simple linear regression problem.
   Equation is ,
   $y = b_o + b_1x$
   Where,
   y - dependent variable
   x - independent variable
   b_o - Intercept
   b_1 - slope

Multiple Linear Regression:
Here there will be multiple independent variables,
$y = b_o + b_1x + b_2x + b_3x + ….$
The idea here is to find the best fit line equation that can predict the values based on independent variables. Best fit line means the error between the predicted and actual values are very less. So we need to know how well the line fits the data, that is where the cost

function concept comes in. It measures the error between the predicted and actual values. By minimizing the cost function , we will get the most ideal parameter that result the minimum prediction error.

Mean Squared Error is a most common cost function used in linear regression. There are techniques like OLS (Ordinary Least Squares) to minimize the cost function.

The common steps involved in model building using linear regression is ,

1. Data collection
2. Data Preparation
3. Exploratory Data Analysis
4. Train and Test data split and processing
5. Scaling
6. Feature selection
7. Model Building
8. Model Evaluation
9. Making predictions

1. Explain the Anscombe's quartet in detail.
   What is Pearson's R? (3 marks)
   Ans:
   Pearson's R is a way of measuring linear correlation between two variables. The value between -1 and 1 indicates the strength and direction of relation. It can be used to summarize the characteristics of a dataset.

Value	Interpretation
0-1	Positive relation(>0.5 - strong), when one changes the other also changes in the same direction
0	No relation
0-(-1)	Negative relation(<-0.5 strong) . When one changes the other changes in opposite direction

2. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

   Ans:

Scaling is a data preprocessing technique which will scale the numerical data to a common scale.

It is important to scale the data to,

- Improve the algorithm performance.
- Prevent the dominance of larger values in numerical feature
- Have a fair comparability between the data

Normalized Scaling : Scales the data between ranges 0-1 typically. It is useful when min and max values are known and distribution is not gaussian

Standardised Scaling : This one scales the data between mean 0 and std deviation of 1. This is ideal when distribution is gaussian and there are outliers present in the data.

3. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans:
Variance Inflation Factor is an indication of multicollinearity between independent variables . An infinite value indicates the high multicollinearity. That means the independent variable has a linear relationship with other independent variables.

$VIF = 1/1-R^2$
Where R^2 - coefficient of determination, this explains the variability of the depended variable

When there is a perfect linear relation this coefficient will be 1 and the VIF will become infinity.

4. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans:
Quantile- Quantile plot is typically used to check the normality of residuals in linear regression. It is a tool that can be used to check if the dataset follows a given distribution.
In linear regression , one of the assumption is residuals are normally distributed. This can be validated using the Q-Q plot. If the points are closely following the straight line , then residuals are normally distributed. We can infer the skewness and potential outliers from the plot.