Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans.

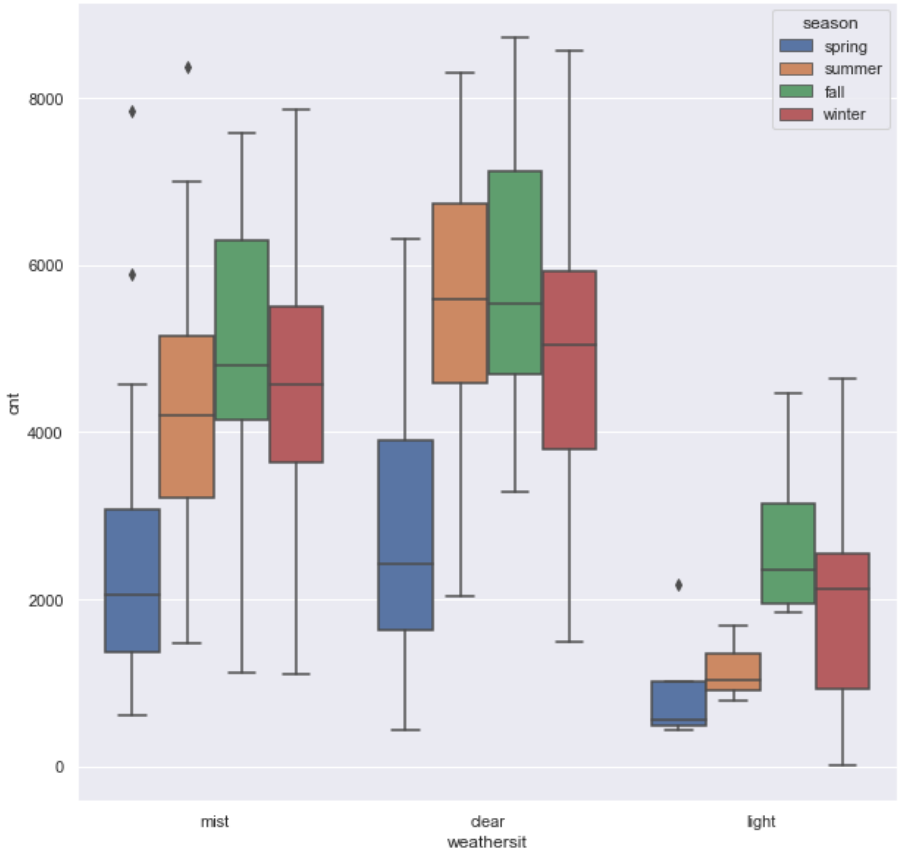

Figure 1: Weather against count, coloured by Season

- The median of bike bookings is cumulatively lowest when light snow, light rain, scattered clouds, or thunderstorm are observed or expected.
- Especially in the season of spring the bike booking median and max are lowest in all weather.
- We can see clear weather has the most bookings. Whereas misty weather has more bookings than when light rain/snow is expected.
- We can infer from this data that the weather and season effect the dependent variable heavily.
- A hypothesis is that in the spring season people might prefer walking.

Q2. Why is it important to use drop_first=True during dummy variable creation?

Ans. During dummy variable creation, the categorical variable with n unique values gets converted into n columns each with a Boolean field. But to represent n values, we only need n-1 columns at max.

To represent A, B, AB and O blood groups, we only need two columns with the following values:

For blood group	Columns	
	A	B
A	1	0
B	0	1
AB	1	1
O	0	0

If we keep all the columns, we can run into collinearity problem as some columns will be redundant. This is a bigger issue when there are not a lot of features. Hence it is recommended to drop first when performing dummy variable creation.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans. Predictor variables **temp** and **atemp** had the highest correlation (0.63) with the target variable.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans. I performed residual analysis after model building to validate the assumptions of Linear Regression.

- Plot the error for each data point in a Seaborn distplot (histogram + distribution line).
- From the plot we can see that the errors are in a normal distribution. This is an assumption of Linear Regression.
- We can also see from the plot y_pred vs y_true that the error terms are evenly distributed around the 0 line which means the error terms are independent and have constant variance.

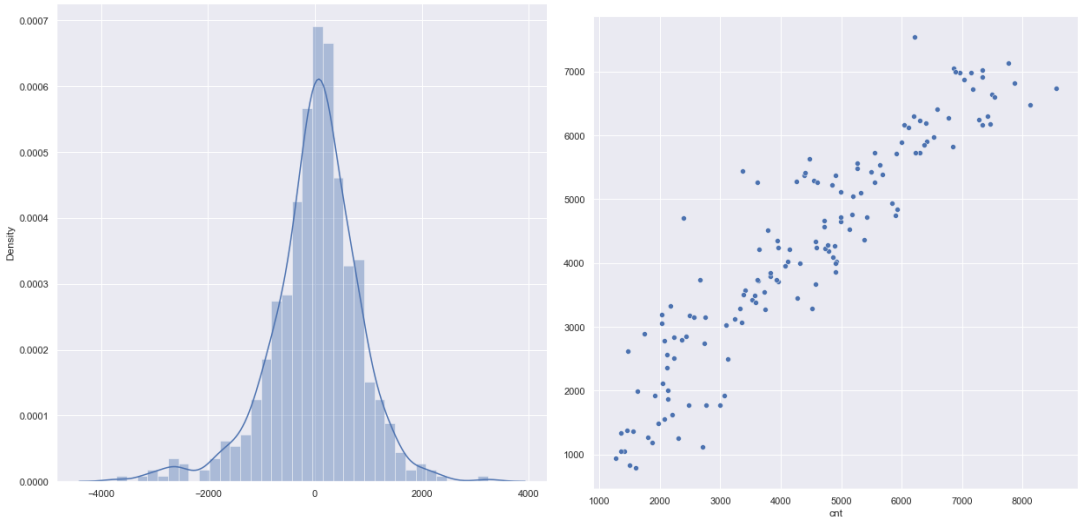

Figure 2: A: Distplot of error terms B: Y pred vs Y true

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans. The following features contribute significantly towards explaining the demand of the shared bikes:

1. Temp
2. Year
3. Weather_light

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Ans. Linear regression is a statistical method to find the calculate a relation between independent variables and dependent variables to predict the dependent variable given the independent variable. Linear regression assumes that there is a linear relation between the dependent variable and independent variables. Linear regression does this by fitting a line/hyperplane through the data which best explains the variance in the dependent variables given the dependent variable. The formula for Simple linear regression is:

$y = \beta_0 + \beta_1 x$ where y is the dependent variable, β_0 is the zero intercept of the straight line and β_1 is the slope of the straight line.

The algorithm learns , β_0 and β_1 by:

1. Random initialization
2. Calculate cost function, our aim is to minimize the cost function. For linear regression the cost function is usually MSE (mean squared error)
3. Perform gradient descent to find delta in the coefficients.
4. Reduce delta from the coefficient by a factor of learning rate. Learning rate is a hyperparamter which controls how much we change the coefficients in each iteration.
5. These steps are performed repeatedly till we can no longer reduce the cost function.

The same steps are performed for Multiple linear regression but there are more than 2 coefficients for MLR.

Q2. Explain the Anscombe's quartet in detail.

Ans. - Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.

- Each dataset consists of eleven (x,y) points.
- They were constructed in 1973 by the statistician Francis Anscombe
- The quartet is used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.
- From the graph below, we can see that even if the mean of all datasets is exactly 9 an d variance 11, the data exhibits widely different patterns visually.

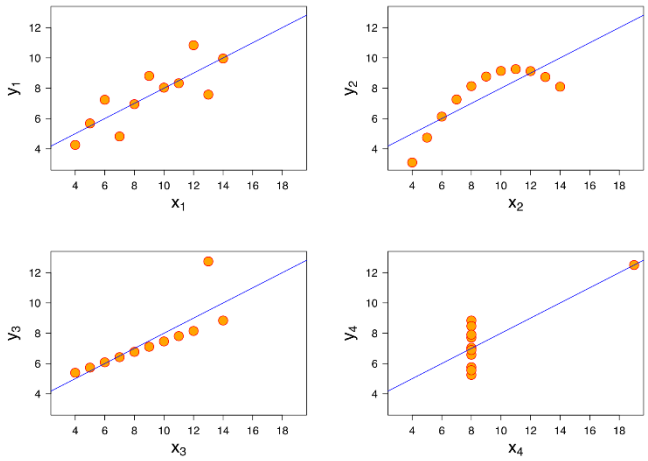

Figure 3: Anscombe's quartet

Q3. What is Pearson's R?

Ans. In statistics, Pearson's R is known as the correlation coefficient.

- It is a measure of 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.
- The Pearson correlation coefficient is symmetric: corr(X,Y) = corr(Y,X) and invariant under scaling eg. aX or bY.
- Formula for Pearson R:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

  -
- Pearson score is directly related to R-squared where R-squared = (Pearson's R)^2

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans.

- Scaling is a preprocessing step applied to dataset to bring the features of dataset to the same scale.
- For eg. In a dataset if we have price of petrol per litre and weight of vehicle in kgs, these two features will have very different minimum and maximum
- Scale can affect linear regression model by skewing the coefficient. Since the weight of vehicle in kg will be orders of magnitude greater than price of petrol per litre, the coefficient of first feature will be orders of magnitude greater than the coefficient of second feature. This might not be true in terms of the model itself because the target variable could be more dependent on the smaller feature instead of the larger feature.
- Scaling is done by the following methods:
  - Normalization: Here values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.
    - Normalization is preferred when data does not follow gaussian distribution.
    - Normalization handles outliers better than other methods.

- Standardization: Here the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.
    - Standardization is preferred when data follows gaussian distribution

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans. The formula for VIF = 1/(1-R2). The only way we get VIF = inf is when R2 = 1. R2 = 1 means that the data is perfectly correlated. When we have perfectly correlated or very highly correlated features in the dataset, the value of VIF can become infinite. To fix this, we should remove the column causing this correlation and try again.

Q6. 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 helps us to determine if two data sets come from populations with a common distribution. This common distribution can be for eg. Normal, exponential or uniform distribution. In linear regression we use Q-Q plot to:

- Check if training and test set come from same population
- To check if y prediction vs y true follow the same distribution and are centred around zero for residual analysis
- Q-Q plot helps us understand if our linear regression model is valid and assumptions of linear regression are validated