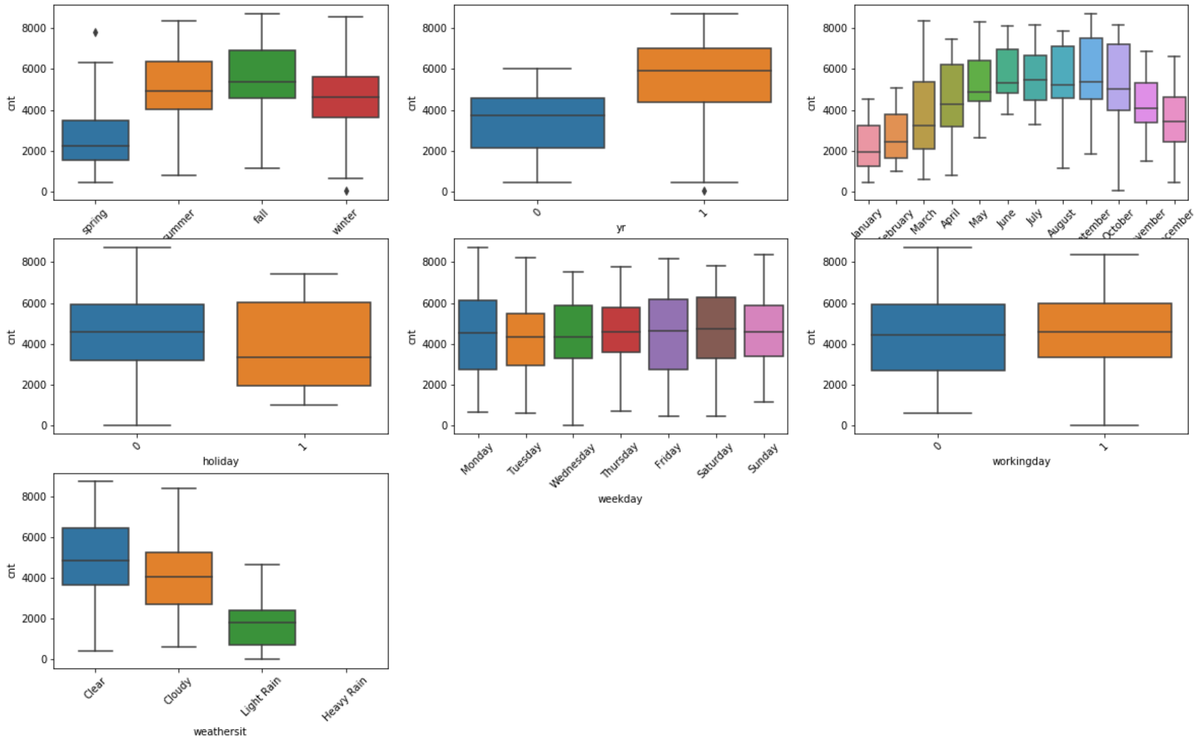
Assignment-based Subjective Questions

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



- For studying the effects of categorical variables on the dependent variable **cnt,** I used boxplots with x axis representing the categorical variables and y axis representing the cnt**.**
- The distribution of the output variable for different categories did provide insights and patterns in the data. I will list my observations below.
- Season:
    - All the quartiles rose from spring to fall. All the quartiles started falling from fall to spring. This could be explained by various weather factors like, rise in temperature, thawing of snow and factors like dry roads.
    - A cyclic pattern is observed for rentals with gradual increase from spring-summer-fall and then a gradual decrease from fall-winter-spring.
    - This is a strong pattern.
- Year:
    - All quartiles rose from 2018 to 2019.
    - As years pass by the popularity for bikes is on the rise.
    - This is a strong pattern, which can be proved if data for more years is available.
- Month:
    - The months from Jan to September show a gradual rise in rentals. The months from October to January show a gradual decline in rentals. This can be explained by the season corresponding to the months.
    - This is cyclic pattern.
    - This is also a strong pattern.
- Holiday:
    - On holidays the IQR is larger but min and max values range is lower.
    - On non-holidays there is a probability of no rentals.
    - This is a weak pattern.
- Weekday:

    - Friday had the lowest rentals for the 1st quartile. Tuesday had the lowest rentals for the 3rd quartile. The IQR is the least for Tuesday, which could imply people did not like to ride bikes on Tuesday for some unknown reason.
    - This is a weak pattern
- Workingday:
    - The 1st quartile had higher rentals for working days. This can be explained as more number of people commute on working days, increasing the number of people that rent bikes, thus increasing the rental count.
    - There is weak pattern.
- Weathersit:
    - During heavy rains the rentals are zero
    - The rentals are very low during light rains and the rentals increase as the weather becomes clear.
    - This a cyclic pattern.
    - There is a strong pattern.

2. Why is it important to use drop_first=True during dummy variable creation?
    - Theoretically, if you have a categorical variable with ‘k’ categories, we need to create ‘k’ different dummy variables.
    - In case the value of categorical variable has a value `v`, then the corresponding dummy variable has the value 1 and the rest of the dummy variable have the value 0.
    - Lets say we have a dummy variable ‘x’ and other dummy variables. If the other dummy variables have a value of 0, we can imply that value of dummy variable ‘x’ would be 1. If one of the other dummy variables has a value of 1, then we can imply that value of dummy variable x would be 0. Thus the value of other dummy variables can help deduce the value of dummy variable x. Thus there is a problem of **multicollinearity** here.
    - Naturally the **VIF** i.e. Variance Inflation Factor for the dummy variable will be high.
    - **Multicollinearity** and **high VIF** should be avoided.
    - **Dropping the dummy variable ‘x’ or in fact any one of the dummy variable removes this multicollinearity.**

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
    - **temp** has 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?
    - First predicted the values of target variable on the training set using the model.
    - Lets say this predictions are stored in y_train_pred.
    - Than we find the difference between actual values of the target variable on the training set i.e y_train and y_train_pred.
    - The difference of values between y_train and y_train_pred is called residuals and can be store in res.
    - We can now plot a **distplot** of the res.
    - If the **distribution** is **normal** and **centered around 0**, our assumption of residuals being normally distributed around 0 is true. **This was the case and the assumption was true in the assignment.**
    - Also we can find patterns between residuals any feature. In this case I used y_train_pred to find patterns between residuals, by plotting a **scatterplot between res and y_train_pred.** If there are **no patterns** we can conclude that the **res are independent. In the assignment this assumption was true.**

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
    - The general approach here is that larger the absolute value of the coefficient, higher is its significance in explaining the target variable
    - The top 3 features for the trained model explaining the demand of the shared bikes are:

Rank	Feature	Coefficient
1	temp	0.518
2	Light Rain	-0.287
3	yr	0.232

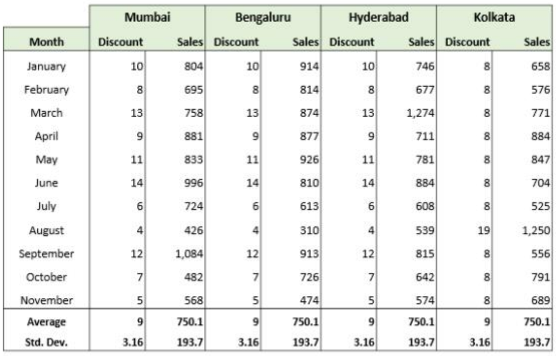
General Subjective Questions

1. Explain the linear regression algorithm in detail.
    - Linear regression is a supervised learning algorithm used when the **target variable is continuous**. It is used to model the linear relationship between the target variables and other features in the dataset.
    - Before building a linear model we establish the assumptions below:
        - We should be able to linearly predict the value of y, using the features X. X can be multiple features and y is a single continuous variable. The equation of the model will be if we have n features in X:
            - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \in$
        - The features comprising X should be independent i.e. they should not exhibit multicollinearity.
        - The error terms, i.e. residuals should be normally distributed around 0.
        - The error terms should be independent
    - General Linear regression algorithm:
        - Split dataset into train and test sets.
        - Select the features i.e. **X** for the model that can determine the target variable **y**.
        - Select a **cost function** which can be reduced to get the best fitting model. For our case we select Ordinary Least Sqaures i.e. OLS.
        - Now **using the cost function** find the best model that fits the train data. By this we mean it learns the β_i coefficients of the linear equation.
        - Once we have the β_i coefficients we create a model equation
            - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \in$
        - We perform residual analysis.
            - Check if residuals are normally distributed around 0.
            - Check if residuals are independent.
        - We then evaluate the model using the test set.
            - Predict y_{pred} for test set.
            - Calculate R^2 for the test set and compare with the R^2 of the train set. If R^2 values are similar we have a model which generalises well with the test set.
    - The detailed linear regression algorithm has the following **steps**:
        - Read the dataset and understand it.
            - Find numerical and categorical variables

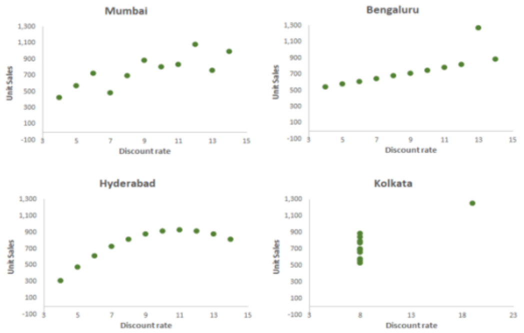
- Prepare the dataset.
    - Convert categorical variables to numeric variables.
        - Convert binary variables to 1/0.
        - Convert non binary variables to dummy variables and drop the original binary variables.
    - Scale the numeric variables.
    - Split dataset into train and test
    - Split each set into y and X.
- Include all independent features in X.
- Build the model on the training set.
    - Build model.
    - Drop features with high pvalue or drop features with high VIF.
    - If a feature is dropped go back to step 1.
    - Now we have the best features and their coefficients at this point.
- Check assumptions on error terms are valid
    - Calculate residuals on the train set.
    - Check if residuals are normally distributed around 0.
    - Check if residuals are independent.
- Analyse on the test set
    - Calculate R^2 for the test set.
    - Compare R^2 of the train and test set.
    - If they are similar or have a very small difference we have model that generalises well with the test set.

2. Explain the Anscombe's quartet in detail.

- Various metrices and their summary might hide the underlying data. Different data may appear to present the same story even though the underlying story may be different.
- The Anscombe's quartet makes this thing clear by providing four data sets with two columns whose summary metrices are very similar to each other to the point that we believe they tell the same story.
- The Anscombe's quartet example is explained with the images below:

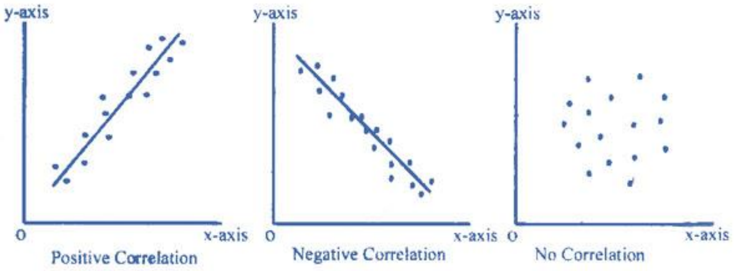


	Mumbai		Bengaluru		Hyderabad		Kolkata	
Month	Discount	Sales	Discount	Sales	Discount	Sales	Discount	Sales
January	10	804	10	914	10	746	8	658
February	8	695	8	814	8	677	8	576
March	13	758	13	874	13	1,274	8	771
April	9	881	9	877	9	711	8	884
May	11	833	11	926	11	781	8	847
June	14	996	14	810	14	884	8	704
July	6	724	6	613	6	608	8	525
August	4	426	4	310	4	539	19	1,250
September	12	1,084	12	913	12	815	8	556
October	7	482	7	726	7	642	8	791
November	5	568	5	474	5	574	8	689
Average	9	750.1	9	750.1	9	750.1	9	750.1
Std. Dev.	3.16	193.7	3.16	193.7	3.16	193.7	3.16	193.7



- As we can see the table readings show that the summary metrices are similar, giving us an impression that the data is very similar.
- When we actually plot the data it becomes clear to us that it is not the case. The actual distribution of data is very different for the four data sets.
- Thus the intention of Anscombe's quartet is to stress the importance of data visualisation.

3. What is Pearson’s R?
    - Pearson’s R is also know as the correlation between two variables. It is a metric that measures linear correlation between two variables X and Y.
    - It can have values between -1 to +1
    - -1 indicates a strong negative relation between X and Y i.e. as X increases Y decreases.
    - +1 indicates a strong positive relation between X and Y i.e. as X increases Y increases.
    - When value is close to 0 or is 0, the relation between X and Y becomes weak or 0, i.e. change in X does not affect Y.
    - The image below explains this:



Positive Correlation Negative Correlation No Correlation

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
    - Scaling is a transform applied to values to enlarge or minimize its representation such that its intent/proportion is not lost.
    - If feature scaling is not performed in machine learning the algorithm gives **higher weightage to variables with larger scales and lower weightage to variables with lower scales**. This can be avoided by scaling the features to same levels.
    - Another reason for scaling data is that some **machine learning algorithms converge much faster when features are scaled**.
    - Normalising a feature scales the feature to values between 0 to 1. This scaling gets the difference between a data point and the min value and calculates its proportion to difference between max and min points.
        - If x_{min} and x_{max} are min and max values, the normalised value of a data point x_i is given by the formula: $x_{ni} = x_i - x_{min} / x_{max} - x_{min}$
    - Standardising a feature scales the distribution to a standard normal distribution with mean zero and standard deviation of one.
        - The standardised value of a value x_i, and a mean of x given by mean(x) and standard deviation given by sd(x) is given by the formula: $x_{si} = x_i – \text{mean}(x) / \text{sd}(x)$

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
    - R^2 how well a linear model fits the dataset. A high value indicates that the large value of variance of the target variable is explained by the independent variables.
    - The formula for VIF: $1/(1 – R^2)$
    - It is clear that if R^2 increases VIF increase. If R^2 becomes 1, VIF will be infinite.
    - An R^2 value of 1 indicates a very good explanation of variance of target variable by the independent variables.
    - If an independent variable is very well explained by other independent variables then its R^2 will be 1 and hence the VIF will be ∞.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
    - Quantile-Quantile plots are also known as Q-Q plots. It lets us find out the type of distribution that matches the given variable if we take a sample with a small sample size.
    - For a large sample size if we plot the distribution it is easy to deduce the type of distribution of the population variable.
    - If we have a small sample size it is not possible to deduce the type of distribution of the population.
    - The X axis plots the actual sample data, while Y axis plots the predicted values assuming the distribution type that we want. In most cases we want the to prove that the population variable has a normal plot.
    - If the plot reveals a solid line, then we can definitely say that the distribution of the population variables is of the assumed distribution i.e. normal distribution.
    - If the plot does not reveal a straight line then the population variable follows a skewed distribution.