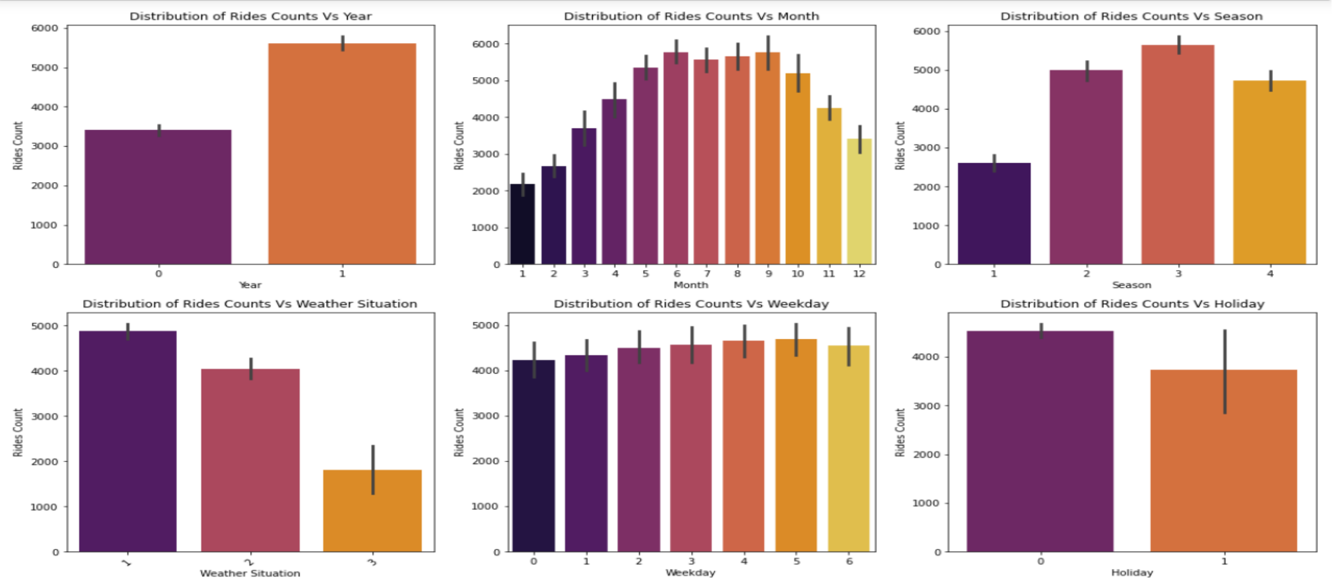
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

The categorical variables identified on the dataset are yr, mnth, season, weathersit, weekday, holiday and workingday.



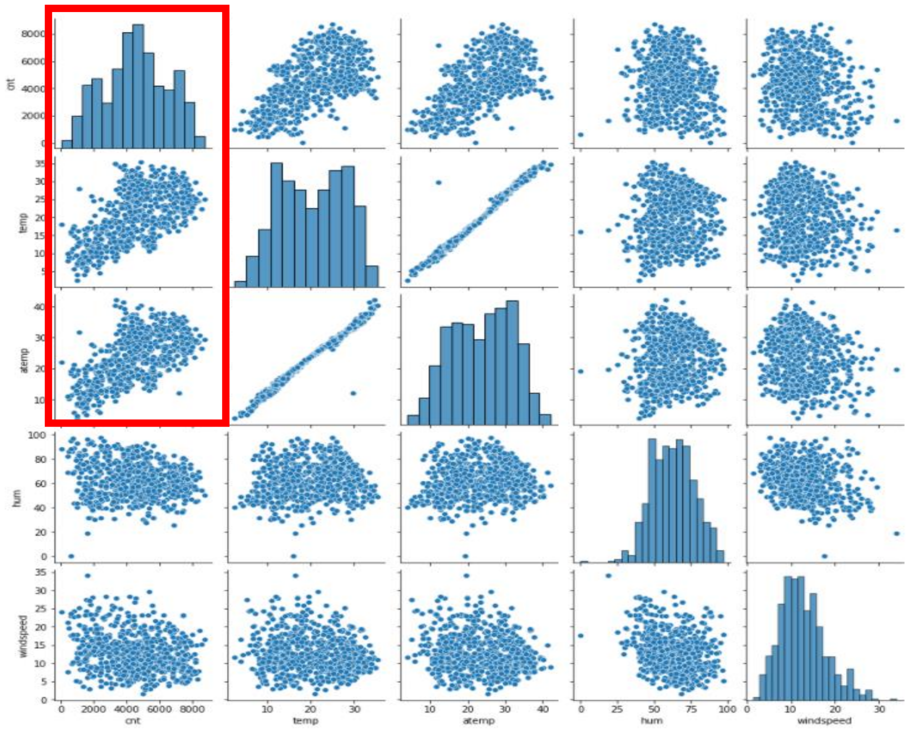
Year - As we can see in the above plot, there is a good jump in the number of rides from 2018 to 2019.
Month - There is definitely some pattern in the distribution of ride counts across the months. The ride counts are higher during the middle months(May-Oct) and relatively low in the year beginning and at the end.
Season - Ride count is more during the seasons Fall, Summer and Winter seasons.
Weathersit - Majority rides are taken when the weather is Clear, Few Clouds, Partly cloudy and it is least during heavy rain/snow.
Weekday - The ride counts are close across all the days of the week.
Holiday - Ride counts are lower during the holidays.
Workingday - The ride counts are little higher on working days.

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

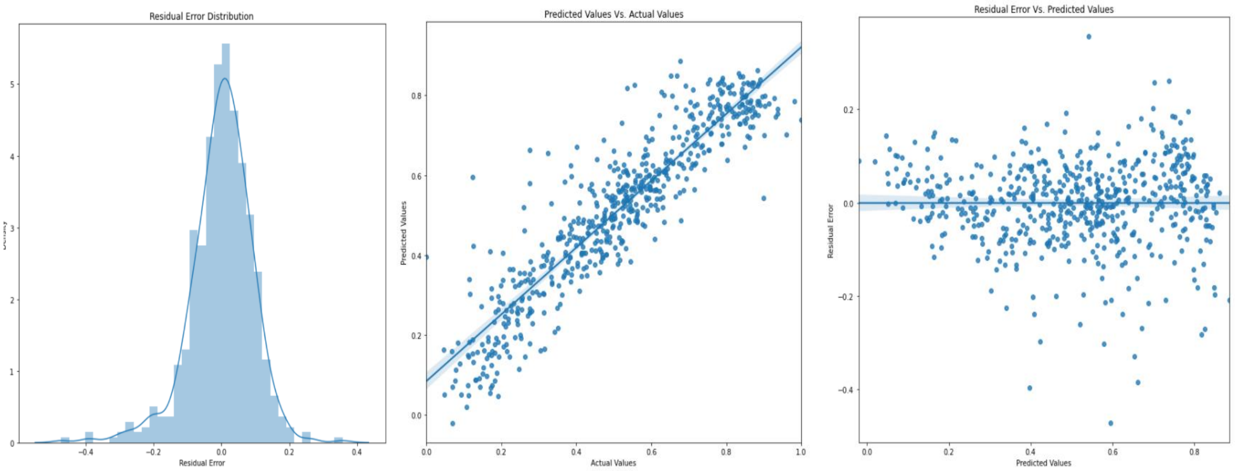
- The first column will be redundant, and all the other dummy variables can explain this variable and they are highly correlated i.e., there will be multicollinearity between dummy variables.
- The VIF score will be infinity for this feature/column (because R-squared for this variable will be 1, as other variables will explain it completely).
- Even though we are proceeding with the model creation with this variable, we will eventually drop this during feature reduction (considering its high VIF score).
- But having this variable can distort the significance of the variables and makes it harder to choose.
- Avoiding/not having Multicollinearity is one of the prime assumptions of Multiple Linear Regression. Hence, we to need to drop the first column after performing **One Hot encoding**.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

As seen below, ‘**temp**’ has the highest correlation with target variable ‘cnt’. As ‘atemp’ is highly correlated with ‘temp’, both ‘temp’ and ‘atemp’ are highly correlated with ‘cnt’.



4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)



a) **Error Terms are normally distributed** – As seen in the plot 1, plotting a distribution plot for the residual errors can help us confirm this. The error terms are normally distributed with mean centered around 0. (*residual_error = y_train - y_train_pred)*

sns.distplot(residual_error)

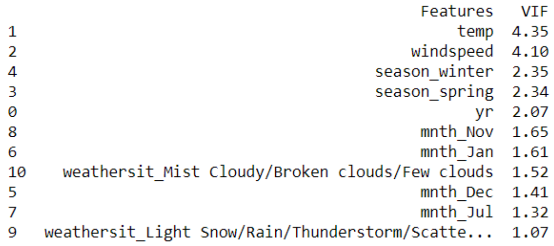
b) **Homoscedasticity** (Probability distribution of errors has constant variance) – As seen in plot 2, plotting a regplot between y_train and y_train_pred can help us confirm this.

We can see that residuals are equally distributed across predicted value - which confirms Homoscedasticity This means we see equal variance of the error terms across the dataset and it is not concentrated in only few places. *sns.regplot(x=y_train, y=y_train_pred)*

c) **Error Terms are independent** – As seen in plot 3, plotting a regplot between residual_error and y_train_pred can help us confirm this. Clearly, all the residuals are scattered across the plot and don't follow any pattern. This confirms that the error terms are independent.

sns.regplot(y=residual_error, x=y_train_pred)

d) **Multicollinearity –** As seen in the below snapshot, calculating the VIF scores between the features of the model confirm this. Clearly the VIF scores for the features are less than 5, which explains that the features are not closely associated.



```
                                            Features   VIF
1                                               temp  4.35
2                                          windspeed  4.10
4                                      season_winter  2.35
3                                      season_spring  2.34
0                                                 yr  2.07
8                                           mnth_Nov  1.65
6                                           mnth_Jan  1.61
10     weathersit_Mist Cloudy/Broken clouds/Few clouds  1.52
5                                           mnth_Dec  1.41
7                                           mnth_Jul  1.32
9   weathersit_Light Snow/Rain/Thunderstorm/Scatte...  1.07
```

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

The top 3 features contributing significantly towards explaining the demand of the shared bikes are:

```
Temp: 0.419286
weathersit_Light Snow/Rain/Thunderstorm/Scattered clouds: -0.291622
Yr: 0.232241
```

1. Temperature has the highest positive coefficient, which indicates that the higher temperatures is what riders prefer the most.
2. Year on year, number of riders who are taking the rides are increasing.
3. Riders doesn't prefer to ride during rains and thunderstorms.

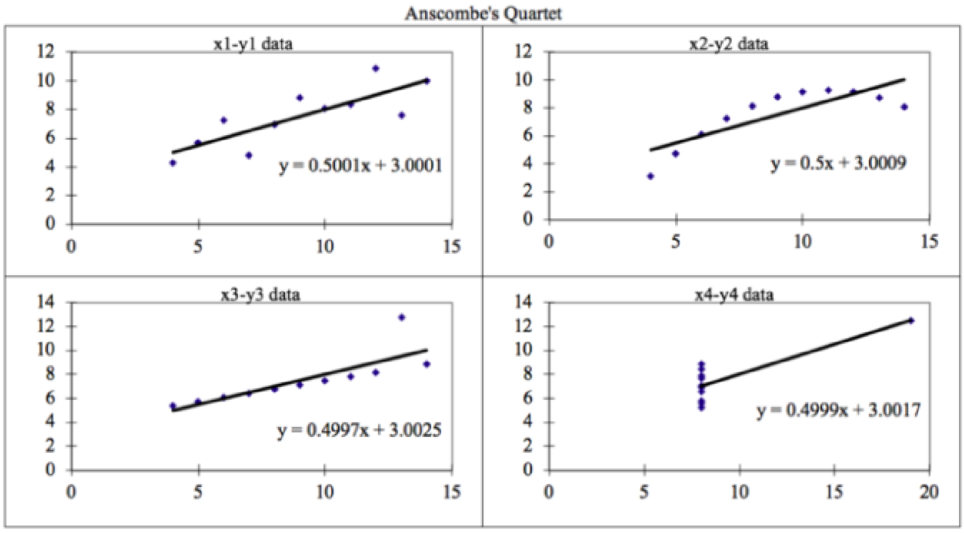
General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
    - Linear regression is a supervised linear approach/ statistical model that identifies the relationship between a target variable (dependent feature) and one or more explanatory variables (independent features).
    - The relationship between the variables is identified by fitting a line/linear equation to observed data.
    - A linear regression line takes the form **y = mx + c** , where 'x' is the explanatory/independent variable and 'y' is the dependent variable. 'm' is regression coefficient and 'c' is constant.
    - Linear regression is broadly classified into 2 types.
        - ➢ *Simple Linear Regression(SLR)*: Here we predict one dependent variable using one independent variable. It takes the form y = mx + c
        - ➢ *Multiple Linear Regression(MLR)*: Here we predict one dependent variable using more than one independent variables. It takes the form y = a0 + a1x1 + a2x2 + …
          Where x1, x2,.. are independent variables. a0, a1, a2,.. are the coefficients of the variables that will be predicted by the model

- Linear regression model is built with an aim to identify the coefficients which has the least error. This is usually done by either ‘differentiating the variables and equating it to zero’ or ‘using Gradient Descent approach’.
- The dependent variable here should be continuous variable and independent variables can be either continuous or categorical variables.
- Linear regression is most basic form of regression analysis. Three major uses for regression analysis are
    - Identifying the strength of predictors – which factors influence the dependent variable
    - Forecasting an effect – what could be the probable value for a variable
    - Trend forecasting – identifying how the trend for a given variable will be

2. Explain the Anscombe’s quartet in detail. (3 marks)

- Anscombe’s quartet explains the importance of plotting the graphs before the analysing and building the models, and the effect of other observations on statistical properties.
- This is developed by statistician Francis Anscombe, where he came up with 4 datasets (containing 11 datapoints each). These 4 sets have almost same statistical observations, like mean, variance etc.
- When a regression is built on these datasets, all of them represent a similar model, however the distribution of these points is completely different from each other when graphed.



- First dataset fits the model well.
- Second dataset follows a non-linear pattern and doesn’t fit well to the model.
- Third dataset, though there is a linear relationship, has outliers in the data.
- Fourth dataset, while there are only few data points(high-leverage points) explained by the model, a high correlation coefficient can be produced.

3. What is Pearson’s R? (3 marks)

- Pearson’s R, also known as Pearson’s correlation coefficient, is a measure of linear correlation between two sets of data i.e., to what degree 2 variables are related to each other. It gives you the measure of the strength of association between two variables.
- Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

- The range of the correlation coefficient is from -1 to 1.
  - An r of -1 indicates a perfect negative linear relationship between variables
  - an r of 0 indicates no linear relationship between variables
  - an r of 1 indicates a perfect positive linear relationship between variables
- Pearson's correlation assumption: at least one variable must follow a normal distribution
- For SLR, the coefficient of Determination(R^2) is the square of Coefficient of Correlation(r).

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

- Scaling is a method of standardising the independent variables in the given dataset i.e., bring down all the input/independent features to a common scale/range.
- There are 2 commonly used techniques for scaling.
  - Normalization (Min-Max scaling)
  - Standardization
- The explanatory features can be of different ranges/scales based on their purpose (area, temperature etc.). If features are not scaled, there can be a great difference in the coefficients assigned by the model. The significance of a feature does not come out clear because of this.
- Also, the algorithm tends to weigh greater values higher and smaller values lower, irrespective of their units of measurement.
- Normalized scaling tries to fit the range of input values between 0 and 1.

  $X_{new} = (X_i – \min(X))/(\max(X) – \min(X))$
- Standardization scales the feature to have a distribution with mean equal to 0 and variance equal to 1. The scaled values can be both positive/negative and can have higher values for outliers.

  $X_{new} = (X_i – X_{mean})/(\text{std. deviation})$

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

- Variance Inflation Factor (VIF) is a measure of multicollinearity in regression analysis.
- Presence of correlation between variables (high VIF values) can badly affect the model results. VIF tells how much variance of regression coefficient is being exaggerated by the relationship with the other variables(collinearity).
- VIF for an independent variable is calculated as below

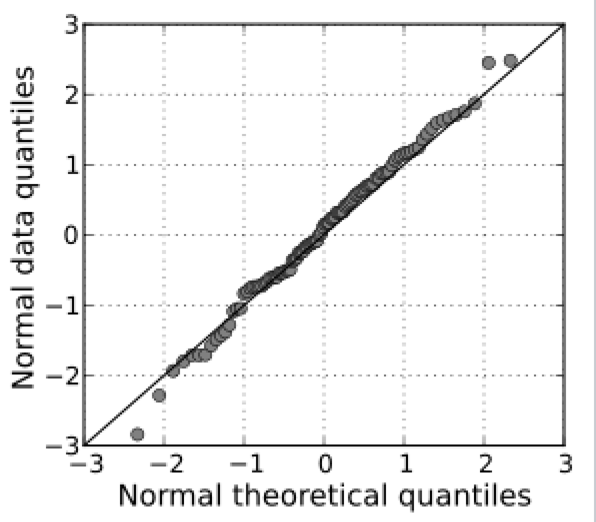
  $VIF_i = 1/(1-R_i^2)$
- When there is a perfect correlation, R^2 value will be 1, that is, given independent variable is perfectly explained by other independent variables in the dataset, the VIF = **infinity**

  This usually happens **when we do not drop the first column while creating dummy variables**.
- It is preferred to drop the independent variables with high VIF scores for better accuracy of the model and to satisfy the primary assumption of linear regression model that the explanatory variables should not be correlated.

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

(3 marks)

- A Q-Q plot is a graphical tool which assists us to assess if a given set of data plausibly follows a certain theoretical distribution or not(like normal or exponential distribution). Basically, this plot is used to determine if two datasets came from populations with a common distribution(shape).
- It is a scatterplot created by plotting two sets of quantiles(percentiles) against each other. If both the datasets represent the same theoretical distribution, then we can see the plotted points forming a line.
- Use and Importance of Q-Q plot in linear regression:
  - ➢ Can help to validate if the train and test datasets came from populations with same distributions.
  - ➢ Can assist on verifying if both sets of data have similar distributional shapes (normal/exponential etc.) and their tail behaviour.
  - ➢ Can be used to check if the datasets have common location and scale.


Normal data quantiles
Normal theoretical quantiles