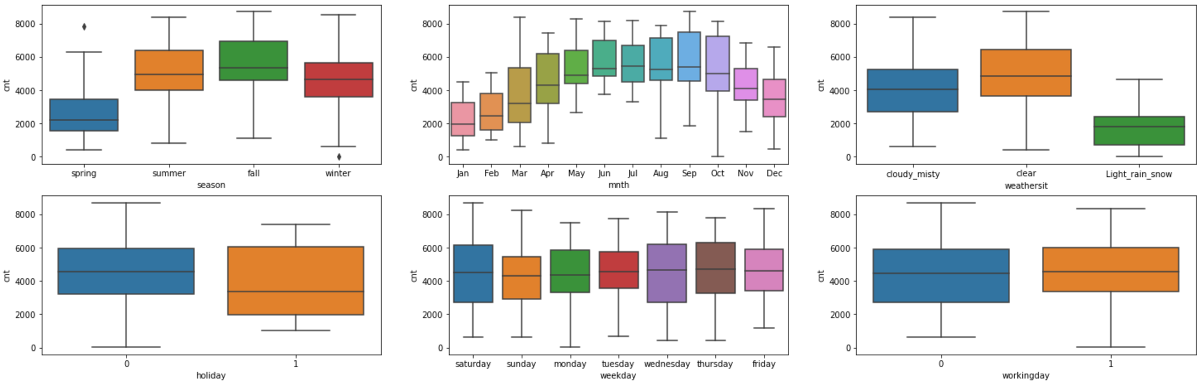
Assignment-based Subjective Questions

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

   Season, Weekday, Month and Weather are the categorical variables in this dataset.



- As we can see from the above bar plots, Summer and Fall season sees an increase in bike sharing, with its respective months May to Sept seeing a similar trend.
- As expected, rain and snow hinder bike riding while a clear and slightly cloudy weather is best suited.
- No particular weekday has an increased effect on the bike shares, with all of them having a similar range.

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

   When creating dummy variables, if there are n category levels, then n dummy variables are created. But this can easily be achieved with n-1 variables.

   For example: Let's assume a weather variable which has 3 levels – Clear, Mist, Rain

   When created without drop_first:

Clear	Mist	Rain
0	0	1
0	1	0
1	0	0

This is redundant because, if the weather is neither misty nor rainy, it can be inferred that it is clear. This can be achieved by dropping the first variable

When created with drop_first:

Mist	Rain
0	1
1	0
0	0

The last row means that the weather is clear.

Hence, it is important to use **drop_first=True**, so that there are no redundant columns.

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

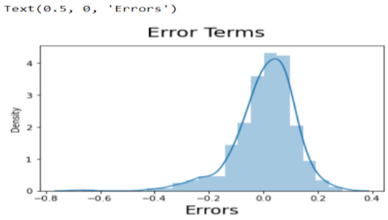
   Temp & atemp (0.64 & 0.65) variables have the most correlation with the cnt variable

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

   The assumptions are:

   1. Linear relationship between the predictor and target variables
   2. Error terms are normally distributed
   3. Error terms are independent of each other
   4. Error terms have constant variance

   On plotting a histogram on the error terms (actual value-predicted value), show that the error terms are normally distributed and that the mean is 0



```
Text(0.5, 0, 'Errors')
```

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

   Temperature, Year and September Month & Winter contribute to increasing the bike demand

   a. We saw an increase in bike demand in 2019 compared to 2018. So, when the pandemic ends, and things return to normal, the business can see a spike in demand
   b. We saw that pleasant temperature contributes to demand
   c. We can see that temperature variable is having the highest coefficient of 0.4612, th at translates to “when temperature increases by one unit the bike demand increas es by 0.4612 units”
   d. We also see that spring season, month of July, holidays and windspeed have a negative coefficient, which means that during this time, there will be that many units of less demand.

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**

   Predictive analysis allows us to use cleaned & prepared data to build models. One of the types of predictive analysis is Linear Regression. It is one of the most widely used models, used for training the model when there is a linear relationship between continuous variables.

   The model is used to make a prediction on a target value based on independent variables.

If there is a single independent variable, it is called a simple linear regression. And when there is more than one independent variable, it is called multiple linear regression. The linear regression model gives a sloped straight line, depicting the relationship within the variables.

Least squares is a statistical method used to determine the best fit line or the regression line by minimizing the sum of squares created by a mathematical function. The “square” here refers to squaring the distance between a data point and the regression line. The line with the minimum value of the sum of square is the best-fit regression line.

Regression Line, y = mx+c where, y is a dependent variable and x is an independent variable, c is the y-Intercept

Model building steps:

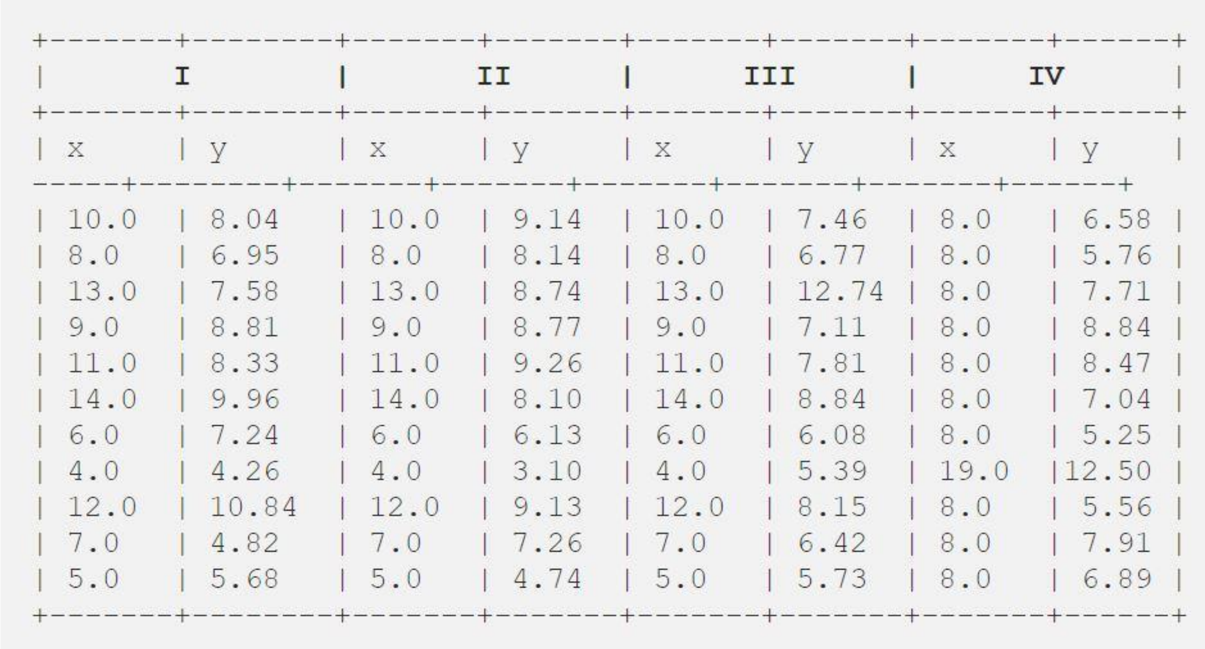
- Create X (predictor variable/s) and y (target variable)
- Create train and test sets (possible 70-30 or 80-20)
- Train the model on the training set (learn the coefficient)
- Residual Analysis
- Predict and Evaluate the model on the test set

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

Anscombe’s quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.
— Wikipedia

Francis John “Frank” Anscombe, a statistician created Anscombe’s Quartet that can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. This illustrates the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x, y points in all four datasets.

Those 4 sets of 11 data-points are given below.



```
+-------+--------+-------+-------+-------+-------+-------+------+
|       I        |      II       |      III      |      IV      |
+-------+--------+-------+-------+-------+-------+-------+------+
| x     | y      | x     | y     | x     | y     | x     | y    |
-----+--------+-------+-------+-------+-------+-------+------+
| 10.0  | 8.04   | 10.0  | 9.14  | 10.0  | 7.46  | 8.0   | 6.58 |
| 8.0   | 6.95   | 8.0   | 8.14  | 8.0   | 6.77  | 8.0   | 5.76 |
| 13.0  | 7.58   | 13.0  | 8.74  | 13.0  | 12.74 | 8.0   | 7.71 |
| 9.0   | 8.81   | 9.0   | 8.77  | 9.0   | 7.11  | 8.0   | 8.84 |
| 11.0  | 8.33   | 11.0  | 9.26  | 11.0  | 7.81  | 8.0   | 8.47 |
| 14.0  | 9.96   | 14.0  | 8.10  | 14.0  | 8.84  | 8.0   | 7.04 |
| 6.0   | 7.24   | 6.0   | 6.13  | 6.0   | 6.08  | 8.0   | 5.25 |
| 4.0   | 4.26   | 4.0   | 3.10  | 4.0   | 5.39  | 19.0  |12.50 |
| 12.0  | 10.84  | 12.0  | 9.13  | 12.0  | 8.15  | 8.0   | 5.56 |
| 7.0   | 4.82   | 7.0   | 7.26  | 7.0   | 6.42  | 8.0   | 7.91 |
| 5.0   | 5.68   | 5.0   | 4.74  | 5.0   | 5.73  | 8.0   | 6.89 |
+-------+--------+-------+-------+-------+-------+-------+------+
```

3. **What is Pearson's R? (3 marks)**

Correlation indicates the measure of how well sets of data are related to each other. The most common measure of correlation is the Pearson Correlation or Pearson's r. It shows the linear relationship between two sets of data. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

Assumptions:

1. Independent of case: Cases should be independent to each other
2. Linear relationship: Two variables should be linearly related to each other. This can be assessed with a scatterplot: plot the value of variables on a scatter diagram, and check if the plot yields a relatively straight line
3. Homoscedasticity: the residuals scatterplot should be roughly rectangular-shaped

Degree of correlation:

1. Perfect: If the value is near ± 1, then it said to be a perfect correlation: as one variable increases, the other variable tends to also increase (if positive) or decrease (if negative)
2. High degree: If the coefficient value lies between ± 0.50 and ± 1, then it is said to be a strong correlation
3. Moderate degree: If the value lies between ± 0.30 and ± 0.49, then it is said to be a medium correlation
4. Low degree: When the value lies below + .29, then it is said to be a small correlation
5. No correlation: When the value is zero

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

   Scaling is used to standardize the independent variables of a data set. During pre-processing of the data, data normalization is done so that all the variables in the same range.

   Example: Let's say that we gave age (18-100), income (5000-20000), experience(0-30), feature scaling would bring these variables in the same range.

   There are 2 techniques to achieve this.

   1. Normalization (Min-Max Scaling)
   2. Standardization

   Normalization

   Also known as min-max scaling or min-max normalization, it is the simplest method and consists of rescaling the range of features to scale the range in [0, 1]. The general formula for normalization is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

   Here, max(x) and min(x) are the maximum and the minimum values of the feature.

   Standardization

   Feature standardization makes the values of each feature in the data have zero mean and unit variance. The general method of calculation is to determine the distribution mean and standard deviation for each feature and calculate the new data point by the following formula:

$$x' = \frac{x - \bar{x}}{\sigma}$$

   Here, σ is the standard deviation of the feature vector, and x̄ is the average.

When to go for Normalization and standardization?

- If the distribution of the quantity is normal, then it should be standardized, otherwise, the data should be normalized. This applies if the range of quantity values is large (10s, 100s, etc.) or small (0.01, 0.0001).
- If you have outliers in your data, they will not be affected by standardization.

There would be no change in t-statistics, f-statistics, p-values, r-squared etc. Only the coefficients are affected after scaling. The choice of using normalization or standardization depends on the problem and the machine learning algorithm

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

   Variance Inflation Factor (VIF) can be used to build a model to explain the predictor using other predictors.

   VIF = $1/(1-R_i^2)$

   R_i^2 is R^2 of the ith variable using other variables, excluding the outcome variable

   A perfect correlation between two independent variables leads to R^2 being 1. When R^2=1, it makes 1/(1-R2) as infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

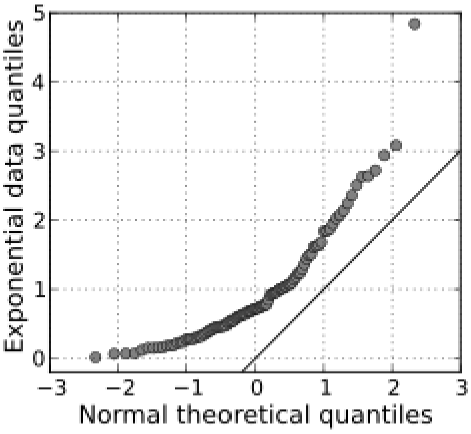
   An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which is shown as infinite VIF).

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

   Quantile-Quantile, (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential, or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution. This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

   Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. A 45 degrees angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

   Below Q Q plot shows the 45 degrees reference line:



A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.

Few advantages:

- It can be used with sample sizes also
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

Q-Q plot can be used in the following scenarios, when two data sets

- Come from populations with a common distribution
- Have common location and scale
- Have similar distributional shapes
- Have similar tail behaviour