## 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)

   **Ans**: In the given dataset, we have season, year, month, weekday, working day, holiday and weather situation columns as categorical variables. As per analysis on these variables, following are the inferences:

   - Season spring has minimal bookings, implies that number of bookings decrease in spring compared to other seasons
   - Year 2019 has a greater number of bookings when compared to previous year, this implies that the business is in good terms.
   - During May month to October month, the bookings are high.
   - If it is a working day, then it seems have more bookings Bivariate analysis.
   - If it's not a holiday, then number of bookings are more as per Bivariate analysis.
   - If the weather is clear, then the number of bookings has a nice count.

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

   **Ans**: the drop_first = True is used for dropping the first occurrence from the list, this is used during the dummy variable creation because, with the help of all the other features given for creating dummy, we can be able to identify which category that row belongs to even if we drop one feature.

   For example: Consider 4 categorical variables, Summer, Spring, Fall and Winter

| Month | Summer | Spring | Fall | Winter |
|---|---|---|---|---|
| 1 | 0 | 0 | 1 | 0 |
| 5 | 1 | 0 | 0 | 0 |
| 11 | 0 | 0 | 0 | 1 |

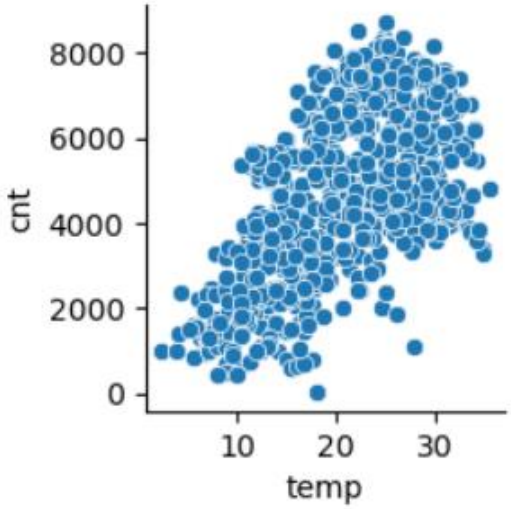
   By dropping the Summer column, the table looks like this

| Month | Spring | Fall | Winter |
|---|---|---|---|
| 1 | 0 | 1 | 0 |
| 5 | 0 | 0 | 0 |
| 11 | 0 | 0 | 1 |

   Without the summer column also, we can identify the second row with month 5 belongs to summer as it doesnot belongs to anyother season. Hence, we can drop this column to minimise our dataframe shape.

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

**Ans**: Temp and atemp have nice trend and highest correlation with the target variable. We can consider any one of the column, Below is the pairplot between temp and cnt column.




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

**Ans**: Validation has been done on the below factors:

- Multicollinearity has been checked and the relationship between the independent variables and the dependant variable is linear.
- Residual Analysis has been done, Error terms are normally distributed.
- Observations are independent of each other.
- Homoscedasticity(Variance of residuals) - The plot at the bottom of the code file shows mostly random scatter, which suggests that the assumption of homoscedasticity is mostly satisfied.

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

**Ans**: Top 3 features:

- Temp has highest positive impact on bike demand
- Light – Light weather conditions, such light snow or light rain has negative impact on bike demand.
- Year variable has high impact, but, here, I am not considering year as a situation like covid can come.
- Winter and September month has positive impact on bike demand

## General Subjective Questions

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

**Ans**: Linear Regression is a form of predictive modelling technique which tells us the relationship between the dependent (target variable) and independent variables. Mathematically, the relationship is represented as follows:

**Y = mX + c or Y = B1X+ B0**

Y – dependent variable / target variable

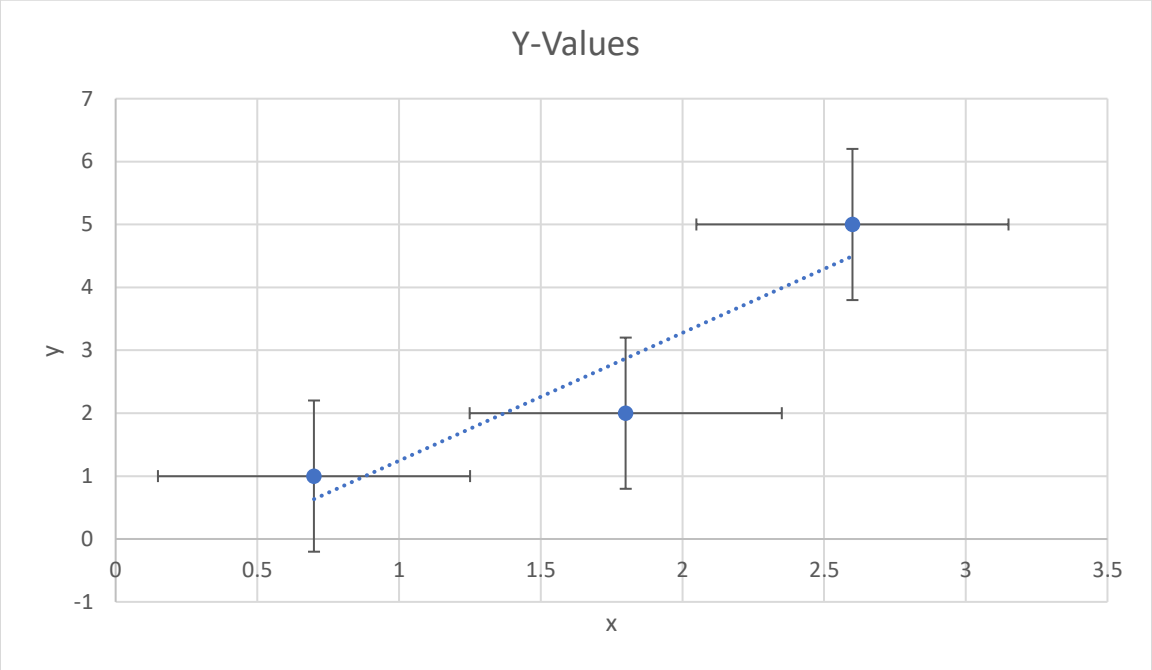
X – independent variable

m – Slope

c – constant, if x=0, Y = c

There are two types of regression:

1. Simple Linear Regression
2. Multiple Linear Regression




Assumptions :

- Linear relationship should be in between X and Y
- Error terms are normally distributed (not X,y)
- Error terms are independent of each other
- Error terms have constant variance

With these assumptions we can go make inferences about the model.

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

**Ans**: Anscombe's quartet, created by Francis Anscombe in 1973, consists of four datasets with 11 (x, y) points each. Here are the key points it emphasizes:

1. **Importance of Visualization**: While numbers like averages, spreads, and relationships (like how things depend on others) are helpful, they might not show the whole story of the data. Looking at pictures of the data is important to see how things are related and any patterns.
2. **Effect of Outliers**: Sometimes, a few points that are far from most of the others can change the averages and the relationships between things a lot. This can make it seem like things are more connected or less connected than they really are.
3. **Validity of Regression Analysis**: When we use methods like drawing lines between points to guess how things are connected, it's important to first look at the data in pictures. If we don't, we might think things are connected one way when they're actually connected differently.

The quartet shows:

- Dataset I: Shows a perfect straight line between points.
- Dataset II: Shows a curved line that a straight line can still be drawn through.
- Dataset III: Has one point far from the others that changes the line and the relationship a lot.
- Dataset IV: Doesn't really have a straight line that fits well, but one outlier makes it seem like it does.

In conclusion, Anscombe's quartet teaches us that looking at pictures of data helps us understand how things are connected and any patterns better than just looking at numbers. Outliers, those points far from the others, can make averages and relationships look different than they really are. This shows why it's important to check if the ways we use numbers to guess relationships actually fit what we see in the pictures.

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

**Ans**: Pearson's R, also known as the Pearson correlation coefficient, is a measure of the strength and direction of the linear relationship between two variables. I have explained in short here:

- **Strength**: Pearson's R tells us how strong the relationship is between two sets of data points. If R is close to +1 or -1, it means the relationship is strong. If it's close to 0, the relationship is weak.
- **Direction**: R also shows the direction of the relationship. If R is positive, it means both variables tend to increase or decrease together (positive correlation). If R is negative, one variable tends to increase while the other decreases (negative correlation).

- **Use**: People use R to understand how two things relate in fields like science, economics, and more. For example, it helps scientists see if there's a connection between two measurements, like how temperature changes with altitude.

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

**Ans**: **Scaling** is the process of adjusting data values to a specific range or distribution. It's performed to ensure variables are comparable and algorithms perform effectively.

**Normalized Scaling** (Min-Max scaling) transforms data to a range typically between 0 and 1. It preserves relative differences between data points and is useful when values need to be within a specific range.

**Standardized Scaling** (Z-score normalization) transforms data to have a mean of 0 and a standard deviation of 1. It centers data around 0, adjusting spread, and is beneficial for algorithms that assume normally distributed data or when features have different scales.

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

**Ans**: Infinite values in the Variance Inflation Factor (VIF) occur when predictors in a regression model are perfectly correlated (perfect multicollinearity). This perfect correlation means one predictor can be precisely predicted from others, leading to Ri^2 being 1 in the VIF formula. So, VIF(Xi) becomes infinite 1/(1-1) indicating severe multicollinearity that indicates the reliability of regression coefficient estimates and model stability.

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

**Ans**: A Q-Q plot (Quantile-Quantile plot) is used in linear regression to check if the residuals follow a normal distribution.

- It compares the quantiles of the residuals against those expected from a normal distribution.
- This helps assess whether the assumption of normality for residuals is reasonable, which is crucial for the validity of regression analysis.
- If the points on the Q-Q plot are pointed out approximately along a straight line, it suggests that the residuals are normally distributed, supporting the reliability of the regression model's results.