Assignment-based Subjective Question

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

Ans:

- Maximum demand noticed in Fall season followed by Summer and Winter. Spring season shows steep decrease in demand.
- Six-month period from May to Oct can be define as high demand period. In Jan the demand is lowest.
- Demand of bike is high on clear day.
- The company shows significant progress in demand from 2018 to 2019.
- Usage of cycle on holiday is slightly lower than the weekdays.
- Usage is similar throughout week irrespective of being working day or not.
- It also shows that the Bikes is also used for office work .

---

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

Ans:

- drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
- Reducing dimensionality with **drop_first=True** improves model efficiency by
  decreasing computational complexity and reducing the risk of overfitting.

---

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

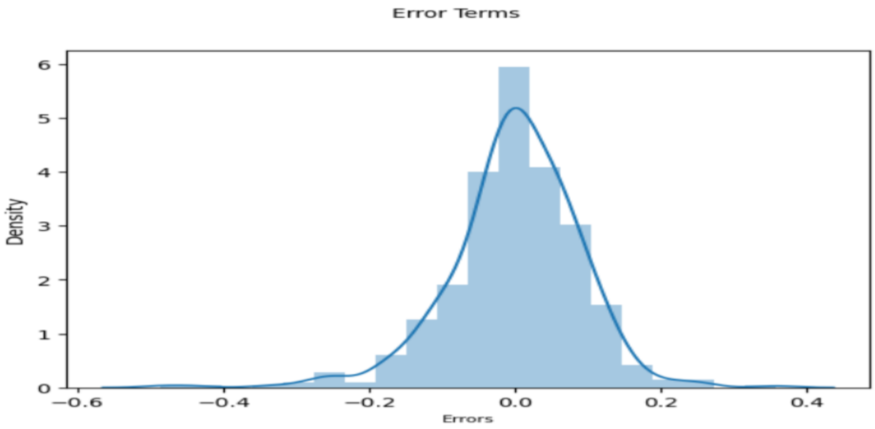
Ans: atemp has the highest correlation with the cnt variable.

---

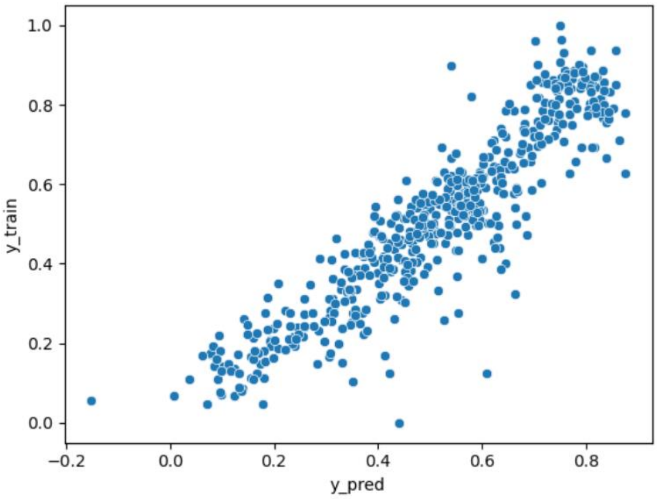
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

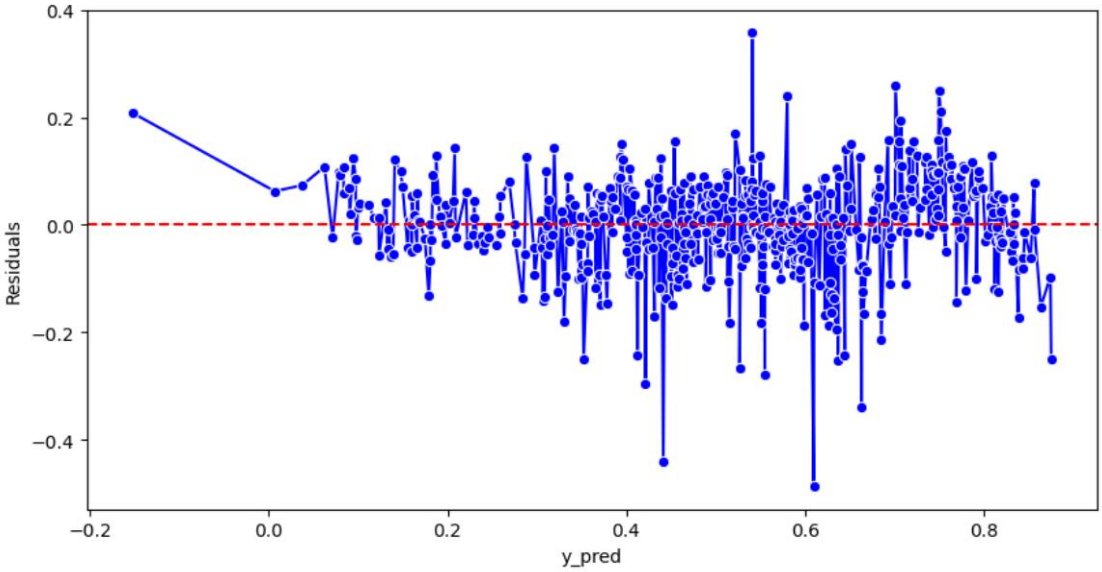
Ans.

- From distplot we can see that, Mean of residual is extremely close to 0. This validates our
  Assumption of normal distribution of error is 0.



- From scatter plot between y_pred and y_train , the plot show an almost constant variance of prediction and thus the error validating the assumption of Homoscedasticity.





- There is no correlation between the error terms.

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

Ans:

- Most important factor is Temperature. With coefficient of 0.3635, for every change in temperature of 1 degree, demand increased by factor of 0.3635.
- The second most important factor is Year with coefficient of 0.2380.
- The third most important factor is windspeed with coefficient of -0.1339. Hence if particular day has strong windspeed the demand will decrease by 13%.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e it finds the linear relationship between the dependent(y) and independent variable(x). In regression we have to find the value of Y, So, a function is required that predicts continuous Y in the case of regression given X as independent features.

Here Y is called a dependent or target variable and X is called an independent variable also known as the predictor of Y. There are many types of functions or modules that can be used for regression.

Linear Equation:

Equation of Simple Linear Regression, where bo is the intercept, b1 is coefficient or slope, x is the independent variable and y is the dependent variable.

$$y = b_o + b_1 x$$

A Linear Regression model's main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized. Error is the difference between the actual value and Predicted value and the goal is to reduce this difference.

Assumptions of Linear Regression –

The basic assumptions of Linear Regression are as follows:

1. **Linearity**: It states that the dependent variable Y should be linearly related to independent variables. This assumption can be checked by plotting a scatter plot between both variables.

2. **Homoscedasticity**: The variance of the error terms should be constant i.e the spread of residuals should be constant for all values of X. This assumption can be checked by plotting a residual plot. If the assumption is violated then the points will form a funnel shape otherwise they will be constant.

3. **Independence/No Multicollinearity**: The variables should be independent of each other i.e no correlation should be there between the independent variables.

4.**Normality**: It assumes that the errors follow a normal distribution.

Evaluation Metrics for Regression Analysis:

1.**R-Squared:** R-squared is a measure of how well the model explains the variance in the dependent variable.

- The formula for R-squared is R-squared=1-(SSR/SST)
- SSR(Sum of Squared Residuals) is the sum of the squared difference between the predicted values and the mean of the actual values.
- SST(Total Sum of Squares) is the sum of the squared difference between the actual values and their mean.

2. **Mean Squared Error (MSE):** Another Common metric for evaluation is Mean squared error which is the mean of the squared difference of actual vs predicted values.

- The formula for MSE is MSE= $(\sum(yi - y^{i})^{2}/n)$

3.**Root Mean Squared Error (RMSE):** It is the root of MSE i.e Root of the mean difference of Actual and Predicted values.

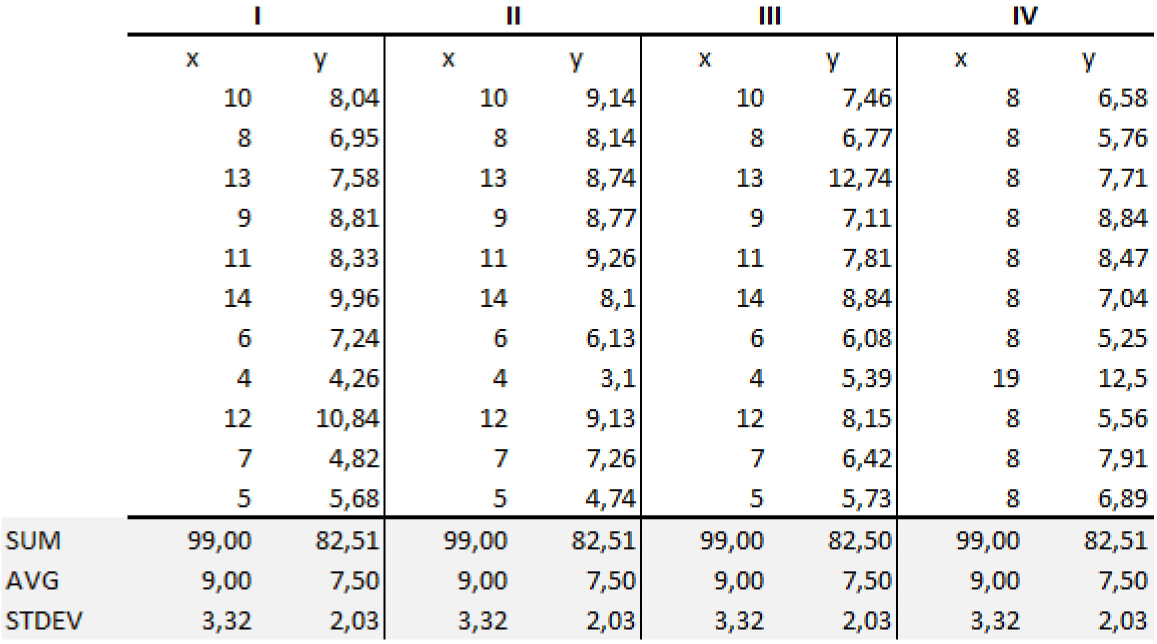
- The formula for MSE is MSE =$\sqrt{(\sum(yi - y\hat{}i)^{2}}\,/n)$

4.**Adjusted R-Squared:** It is improvement to R-squared. It only considers the features which are important for the model and shows the real improvement of the model.

- The formula is $R^2\ Adjusted = 1 - \frac{(1-R^2)(N-1)}{N-p-1}$

2.Explain the Anscombe’s quartet in detail.

Ans: Anscombe’s Quartet was developed by statistician **Francis Anscombe**. This is a method which keeps four datasets, each containing eleven (x, y) pairs. The important thing to note about these datasets is that they share the same descriptive statistics. Each graph tells a different story irrespective of their similar summary statistics. Below is the glimpse of the statistics of the 4 datasets:



	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

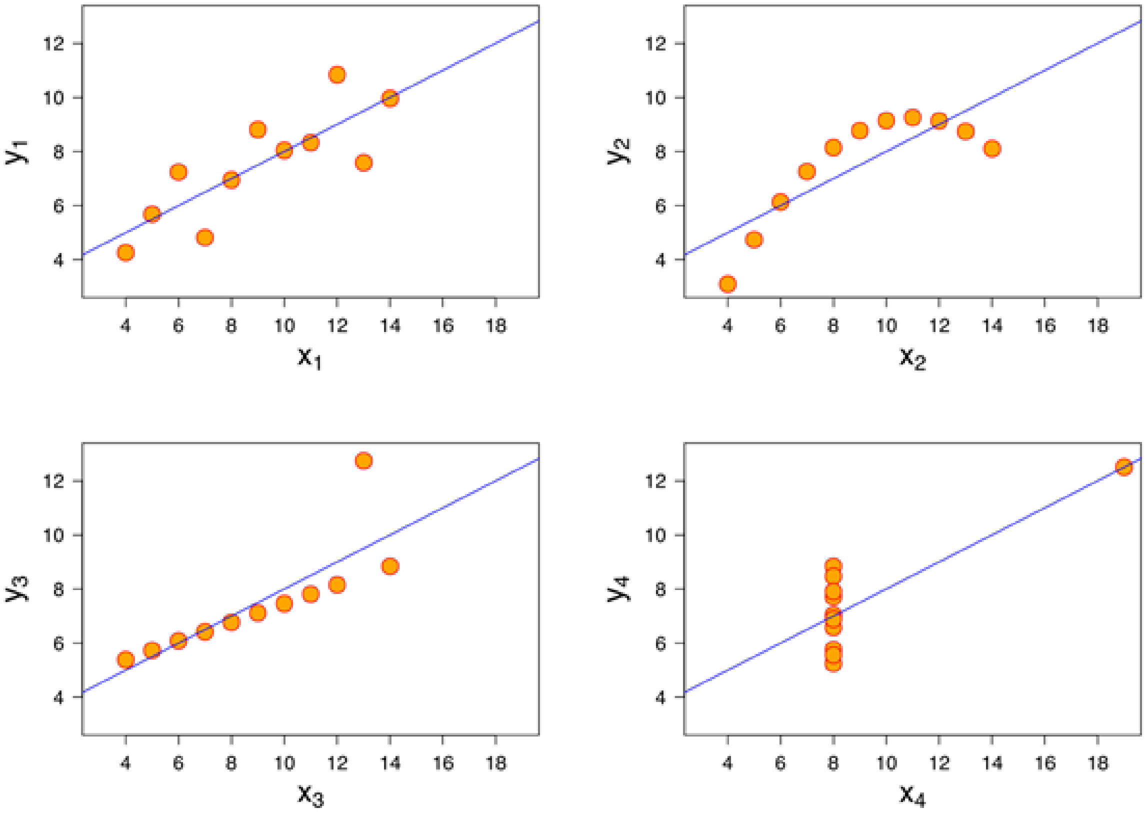
The summary statistics show that the means and the variances were identical for x and y across the groups:

· Mean of x is 9 and mean of y is 7.50 for each dataset.

· Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset

· The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



· Dataset I appears to have clean and well-fitting linear models.

· Dataset II is not distributed normally.

· In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.

· Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

Additionally, Anscombe’s Quartet warns of the dangers of outliers in data sets.

---

3. What is Pearson’s R?

Ans: The Pearson correlation, also called Pearson’s R, is a statical calculation of the strength of two variable relationships. In other words, it’s a measurement of how dependent two variable are on one another. The correlation coefficient formula returns a value between 1 and -1. Here,

- -1 indicates a strong negative relationship
- 1 indicates strong positive relationships
- And a result of zero indicates no relationship at all

Pearson’s Correlation Coefficient Formula

The Pearson’s correlation coefficient formula is the most commonly used and the most popular formula to get the correlation coefficient. It is denoted with the capital “R”. The formula for Pearson’s correlation coefficient is shown below,

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}$$

Here,

n = Number of values or elements

∑x = Sum of 1st values list

∑y = Sum of 2nd values list

∑xy = Sum of the product of 1st and 2nd values

$\sum x^2$ = Sum of squares of 1st values

$\sum y^2$ = Sum of squares of 2nd values

The Pearson’s correlation helps in measuring the strength(it’s given by coefficient r-value between -1 and +1) and the existence (given by p-value)of a linear relationship between the two variables and if the outcome is significant we conclude that the correlation exists.

The interpretation of the Pearson’s correlation coefficient is as follows:-

- A correlation coefficient of 1 means there is a positive increase of a fixed proportion of others, for every positive increase in one variable. Like, the size of the shoe goes up in perfect correlation with foot length.
- If the correlation coefficient is 0, it indicates that there is no relationship between the variables.
- A correlation coefficient of -1 means there is a negative decrease of a fixed proportion, for every positive increase in one variable. Like, the amount of water in a tank will decrease in a perfect correlation with the flow of a water tap.

---

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

Ans: Scaling is the process to normalize the data within a particular range. Many times, in our dataset we see that multiple variables are in different ranges. So, scaling is required to bring them all in a single range.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

The two most discussed scaling methods are **Normalization** and **Standardization**. Normalization typically scales the values into a range of [0,1]. Standardization typically scales data to have a mean of 0 and a standard deviation of 1 (unit variance).

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and 1. **Sklearn.preprocessing.MinMaxScaler** helps to maintain normalization in python.

Formula of Normalized scaling:

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

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean **(μ)** zero and standard deviation one (**σ**).
- **sklearn.preprocessing.scale** helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.

Formula of Standardized scaling:

$$x = \frac{x-mean(x)}{sd(x)}$$

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

Ans: The value of VIF is calculated by the below formula:

$$VIF_i = \frac{1}{1-{R_i}^2}$$

Where, ‘i’ refers to the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

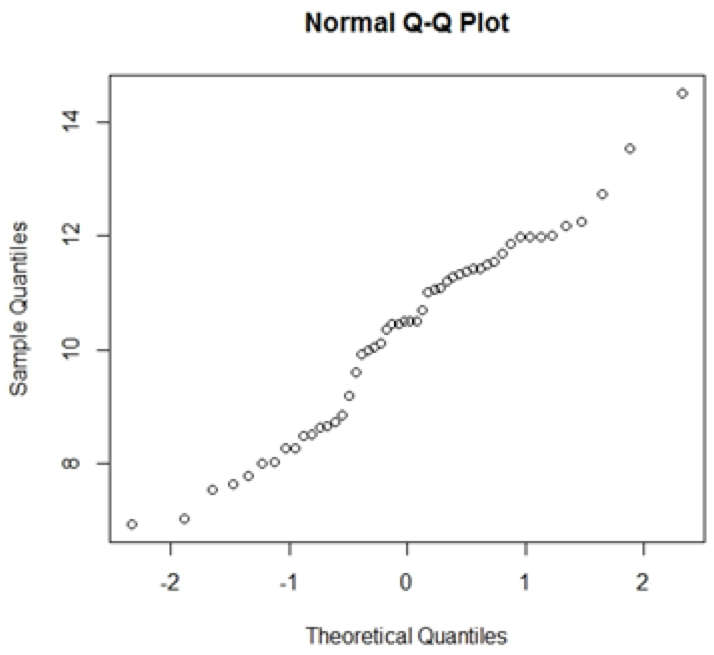
---

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

Ans:

The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that’s roughly straight. Here’s an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



Use of Q-Q plot in Linear Regression: The Q-Q plot is used to see if the points lie approximately on the line. If they don’t, it means, our residuals aren’t Gaussian (Normal) and thus, our errors are also not Gaussian.

Importance of Q-Q plot: Below are the points:

I. The sample sizes do not need to be equal.

II. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.

III. The q-q plot can provide more insight into the nature of the difference than analytical methods.