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?**

Ans: - Below are the observations derived from the analysis of the categorical variables from the dataset

- Season fall has highest demand for rental bikes.
- There is a consistent growth in demand from Jan to June. However, after September, we can see a fall in demand. Also, we can see that, September month is highest in terms of demand. Weather condition can be a major factor here.
- Demand has fall on holidays.
- Demand seems to be high on Friday. However, there is no clarity found about demand basis on weekday data.
- workingday is not giving clear understanding about demand.
- The clear weathersit has highest demand.
- Demand for next year seems to be on higher end.

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

Ans: - drop_first = True is important to use, as it will help in reducing the extra columns created during dummy variable creation.

Syntax: drop_first= bool

Here, default False implies whether to get n-1 dummies from n categorical levels by removing the first level.

When categorical with n levels, idea of dummy variable creation is to build n-1 variables which are indicating the levels.

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

Ans: - Looking at the pair-plot among the numerical variables, '**temp**' variable 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?**

Ans: - I have validated the below 5 assumptions of Linear Regression after building the Model based on training set: -

- Error Terms are normally distributed with mean 0.
- Error terms are independent of each other.
- Error terms have constant variance i.e., homoscedasticity.
- There is insignificant multicollinearity amongst variables.
- There is a linear relationship amongst variables.

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

Ans: - Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes

- temp
- sep
- winter

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Ans: - Linear relationship between variables means that when the value of one or more independent variables will change i.e., increase, or decrease, the value of the dependent variable will also change accordingly.

Mathematically the relationship can be represented with the help of following equation:

$Y = mX + c$

Where, Y is the dependent variable or output variable.

m is slope or gradient of the regression line which represents the effect X has on Y.

c is y value i.e., y- intercept when x=0.

X is the independent variable or predictor variable which is used to make predictions.

The linear relationship can be positive or negative: -

Positive Correlation: - A linear relationship will have positive correlation if both independent and dependent variable increases and if their correlation coefficient has positive value.

Negative Correlation: - A linear relationship will have negative correlation if both independent and dependent variable decreases and if their correlation coefficient has negative value.

Linear regression is of the following two types: -

- Simple Linear Regression

- Multiple Linear Regression.

Assumptions - The following are some assumptions about dataset that is made by Linear Regression model –

- Error Terms should be normally distributed with mean 0.
- Error terms should be independent of each other.
- Error terms should have constant variance i.e., homoscedasticity.
- There should be insignificant multicollinearity amongst variables.
- There should be a linear relationship amongst variables.

Steps: -

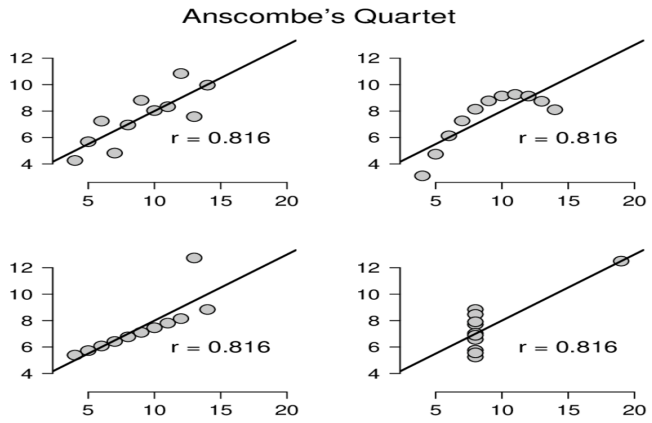
1. Reading and understanding the data
2. Visualizing the data
3. Data preparation
4. Linear Regression Model Building
5. Residual Analysis and Predictions
6. Evaluation and interpretation on the test set.

2. Explain the Anscombe's quartet in detail.

Ans: - Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.

Each dataset consists of eleven (x, y) points. Anscombe's Quartet was developed by statistician Francis Anscombe. These datasets share the same descriptive statistics. But analysis changes, when they are graphed. Each graph represents a different insight irrespective of their similar summary statistics.

Anscombe's quartet highlights the importance of plotting data to confirm the validity of the model fit. In each panel, the Pearson correlation between the x and y values is the same, r =0. 816. In fact, the four different data sets are also equal in terms of the mean and variance of the x and y values.



Anscombe's quartet highlights the importance of plotting data to confirm the validity of the model fit.
In each panel, the Pearson correlation between the x and y values is the same, r = .816. In fact, the four different data sets are also equal in terms of the mean and variance of the x and y values.
Despite the equivalence of the four data patterns in terms of popular summary measures, the graphical displays reveal that the patterns are very different from one another, and that the Pearson correlation (a linear measure of association) is only valid for the data set from the top left panel.
Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models.

3. What is Pearson's R?

Ans: - The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation.

It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Pearson correlation coefficient (*r*)	**Correlation type**	**Interpretation**	**Example**
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the **same direction**.	Baby length & weight: The longer the baby, the heavier their weight.
0	No correlation	There is **no relationship** between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the **opposite direction**.	Elevation & air pressure: The higher the elevation, the lower the air pressure.

The Pearson correlation coefficient (*r*) is the most widely used correlation coefficient and is known by many names:

- Pearson's *r*

- Bivariate correlation
- Pearson product-moment correlation coefficient (PPMCC)
- The correlation coefficient
- The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.
- Although interpretations of the relationship strength vary between disciplines, the table below gives general rules of thumb:

Pearson correlation coefficient (*r*) value	**Strength**	**Direction**
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks) 5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: - Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Features Scaling: -

- Ease of Interpretation

- Faster Convergence for gradient descent methods.

when collected data set contains features highly varying in magnitudes, units and range. In such cases, if scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling.

To solve this issue, we must do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Standardization: -

- It brings all the data into a standard normal distribution with mean 0 and standard deviation 1.
- Mean and standard deviation is used for scaling.
- It is used when we want to ensure zero mean and unit standard deviation.
- It is much less affected by outliers.
- **X= (x-mean(x))/sd(x)**

Minmax Scaling: -

- **It brings all the data in the range of 0 to 1.**
- Minimum and maximum value of features are used for scaling.
- It is used when features are of different scales.
- It is really affected by outliers.
- **X= x-min(x)/(max(x)-min(x)).**

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

Ans: - When there is a perfect correlation, then VIF = infinity.

When the value of VIF is infinite, it represents a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, i.e., VIF= 1/ (1-R2) infinity.

A large value of VIF i.e., Variation Inflation Factor indicates that there is a perfect correlation between the variables.

To fix this, we are supposed to drop one of the variables from the dataset.

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

   Ans: - Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any probability distribution like normal, uniform, exponential.

   Q-Q Plots are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile.

   E.g., the median is a quantile where 50% of the data fall below that point and 50% lie above it.

   Usage: -

   It is also used to compare the shapes of distributions, providing a graphical view by representing similarities and differences between two distributions are similar or different in the two distributions.

Importance: - It Ensures that our ML Model is Based on the Right Distribution. The purpose of Q Q plots is to find out if two sets of data come from the same distribution.