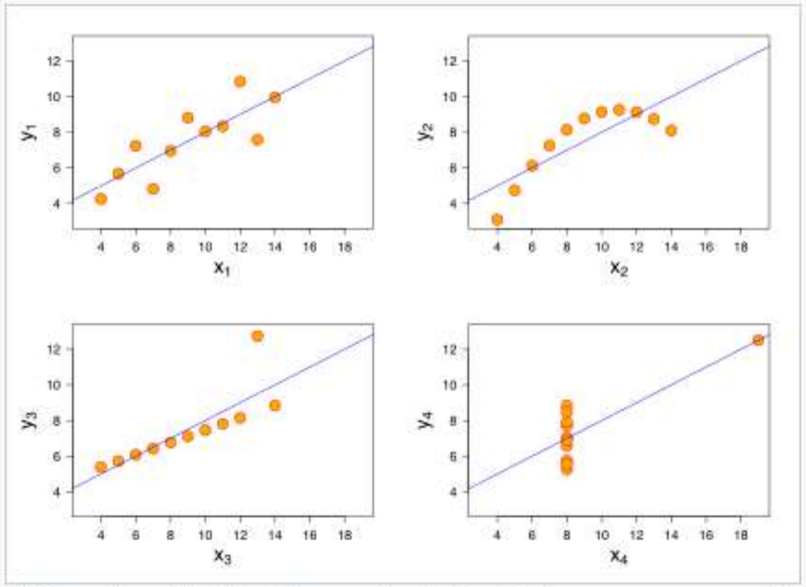
Assignment-based Subjective Questions

Question	**Answer**
From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?	1. The bike hire increases by 0.23 units as an year on year growth 2. The Working Day and Temperature impact the bikes being hired. 3. The seasons of Summer and Winter impact the bikes being hired. 4. The months of August, September and October impact the bikes being hired. 5. The weekday of Saturday impact the bikes being hired. 6. The following climatic / weather condition impact the bikes being hired, NEGATIVELY a. Humidity and Windspeed b. Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist c. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
Why is it important to use drop_first=True during dummy variable creation?	If we do not use **drop_first = True**, then n dummy variables will be created, and these predictors (n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap.
Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?	The **{yr}** and **{temp}** has the highest correlation
How did you validate the assumptions of Linear Regression after building the model on the training set?	Looking at the residual analysis, which has a normal distribution curve and performing a hypothesis testing (Was able to reject the null hypothesis)
Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?	1. The **{yr}** is a significant contributor. 2. The {temp} is a significator contributor 3. The weather condition of high humidity, windspeed and light-snow and light-rain impacts NEGATIVELY.

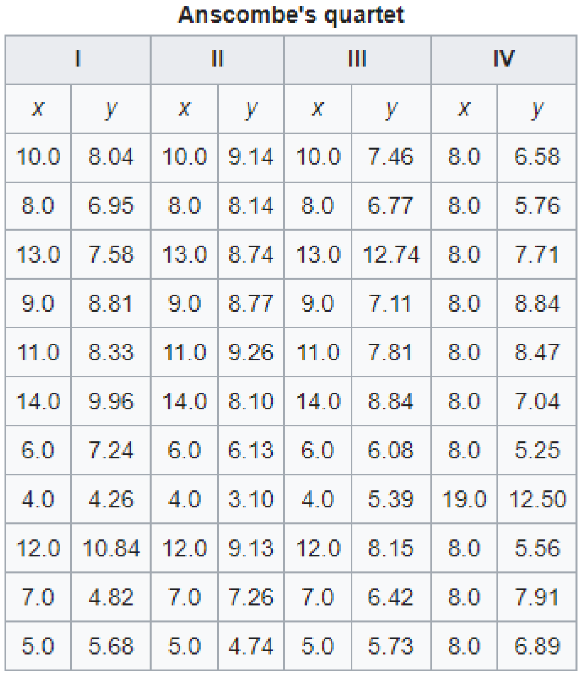
General Subjective Questions

Explain the linear regression algorithm in detail.	The Linear Regression Algorithm has the following steps 1. Import the Data 2. Clean the data by looking for any null values or other data issues 3. Perform the EDA to see correlations and visualize the dependencies among variables in the data. This will also help in earmarking variables that needs to be dropped because of being insignificant. 4. Creating dummy variables for categorical variables 5. Rename the dummy variable columns for being more intuitive while reading 6. Creating the correlation Matrix

	7. Preparing for Train and Test Data sets 8. Scaling the predictors (using Normalization) 9. recursive feature elimination (RFE) taking 25 variables to start with 10. Iterative removal of variables with high p-values 11. Arriving at the optimal model 12. Performing Residual Analysis 13. Performing prediction using the trainer model 14. Comparing values with the actual Test data set 15. Final analysis Report
Explain the Anscombe’s quartet in detail.	**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. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties.  All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.



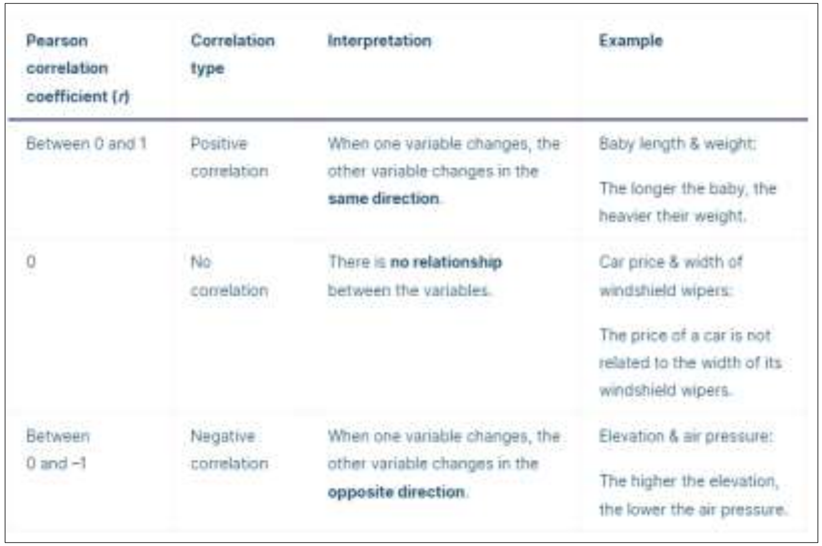
Anscombe's quartet

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

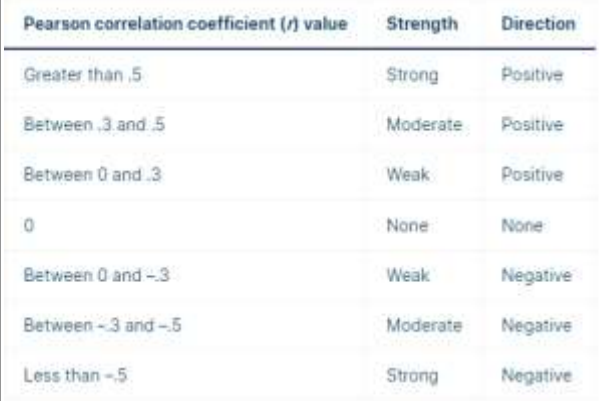
What is Pearson's R?	The Pearson correlation coefficient (r) is the most widely used correlation coefficient and is known by many names:

- Pearson’s r
- Bivariate correlation
- Pearson product-moment correlation coefficient (PPMCC)
- The correlation coefficient

The Pearson correlation coefficient (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.



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

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

Machine learning algorithms like linear regression, logistic regression, neural network, PCA (principal component analysis), etc., that use gradient descent as an optimization technique require data to be scaled. Take a look at the formula for gradient descent below:

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

The presence of feature value X in the formula will affect the step size of the gradient descent. The difference in the ranges of features will cause different step sizes for each feature. To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model.

	Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling. $X^{'} = \frac{X - X_{min}}{X_{max} - X_{min}}$ Here, X_{max} and X_{min} are the maximum and the minimum values of the feature, respectively. • When the value of X is the minimum value in the column, the numerator will be 0, and hence X' is 0 • On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator, and thus the value of X' is 1 • If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1 **Standardization** is another scaling method where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero, and the resultant distribution has a unit standard deviation. $X^{'} = \frac{X - \mu}{\sigma}$ Feature scaling: μ is the mean of the feature values and σ is the standard deviation of the feature values.
You might have observed that sometimes the value of VIF is infinite. Why does this happen?	If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.
What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression	The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

	The advantages of the q-q plot are: • The sample sizes do not need to be equal. • Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.