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?

Answer: season, month, weekday and weathersit are the categorical variables.
For **month**, the months of January and July are correlated and have a mildly significant role and explaining cnt
For **weathersit**, there is a relationship with cnt, there is a siginificant demand when the weather is Light + precipitation or Mist + Cloudy
For **Weekday**, there is a lesser demand on Sunday
For **season**, there is a lesser demand in winter and spring

———————————————————————————————————————

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

Answer: Using drop_first is used so that as it removes the extra columns created when dummy variable are created. Thus it reduces the correlations created among dummy variables.
———————————————————————————————————————

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

Answer: Variable atemp has the highest correlation with the target variable. Correlation coefficient is **0.4452**.
———————————————————————————————————————

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

Answer: We do a 70-30 split between the training and test data. Once results are obtained on the training data, we use the same and validate it against the test data to make sure there is no significant difference between the R-Squared and Adjusted R-Squared values.
———————————————————————————————————————

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

Answer: The 3 features which contribute the most to the demand of shared bikes are atemp, weathersit_Light Precipitation and yr

———————————————————————————————————

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer: Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. When the number of the independent feature, is 1 then it is known as Univariate Linear regression, and in the case of more than one feature, it is known as multivariate linear regression.

The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables. The equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable(s).
In regression set of records are present with X and Y values and these values are used to learn a function so if you want to predict Y from an unknown X this learned function can be used. 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. A linear function is the simplest type of function.

Assumption for Linear Regression Model

Linear regression is a powerful tool for understanding and predicting the behavior of a variable, however, it needs to meet a few conditions in order to be accurate and dependable solutions.

Linearity: The independent and dependent variables have a linear relationship with one another. This implies that changes in the dependent variable follow those in the independent variable(s) in a linear fashion.
Independence: The observations in the dataset are independent of each other. This means that the value of the dependent variable for one

observation does not depend on the value of the dependent variable for another observation.
Homoscedasticity: Across all levels of the independent variable(s), the variance of the errors is constant. This indicates that the amount of the independent variable(s) has no impact on the variance of the errors.
Normality: The errors in the model are normally distributed.
No multicollinearity: There is no high correlation between the independent variables. This indicates that there is little or no correlation between the independent variables.

The equation for linear regression may be described as:

Y = p0 + p1 * x

where,

Y = output variable.
x = input variable. In machine learning, x is the feature, while it is termed the independent variable in statistics. Variable x represents the input information provided to the model at any given time.
p0 = y-axis intercept (or the bias term).
p1 = the regression coefficient or scale factor. In classical statistics, p1 is the equivalent of the slope of the best-fit straight line of the linear regression model.
pi = weights (in general).

The above equation tries to find the line which will best fit the data and provide the best representation which can then be used to predict the output variable Y for a set of values of dependent variables.

Simple and multiple linear regression

The very simplest case of a single scalar predictor variable x and a single scalar response variable y is known as simple linear regression. The extension to multiple and/or vector-valued predictor variables (denoted with a capital X) is known as multiple linear regression, also known as multivariable linear regression.

Multiple linear regression is a generalization of simple linear regression to the case of more than one independent variable, and a special case of general linear models, restricted to one dependent variable. The basic model for multiple linear regression is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \epsilon_i$$

2. Explain the Anscombe’s quartet in detail.

Answer: 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.

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

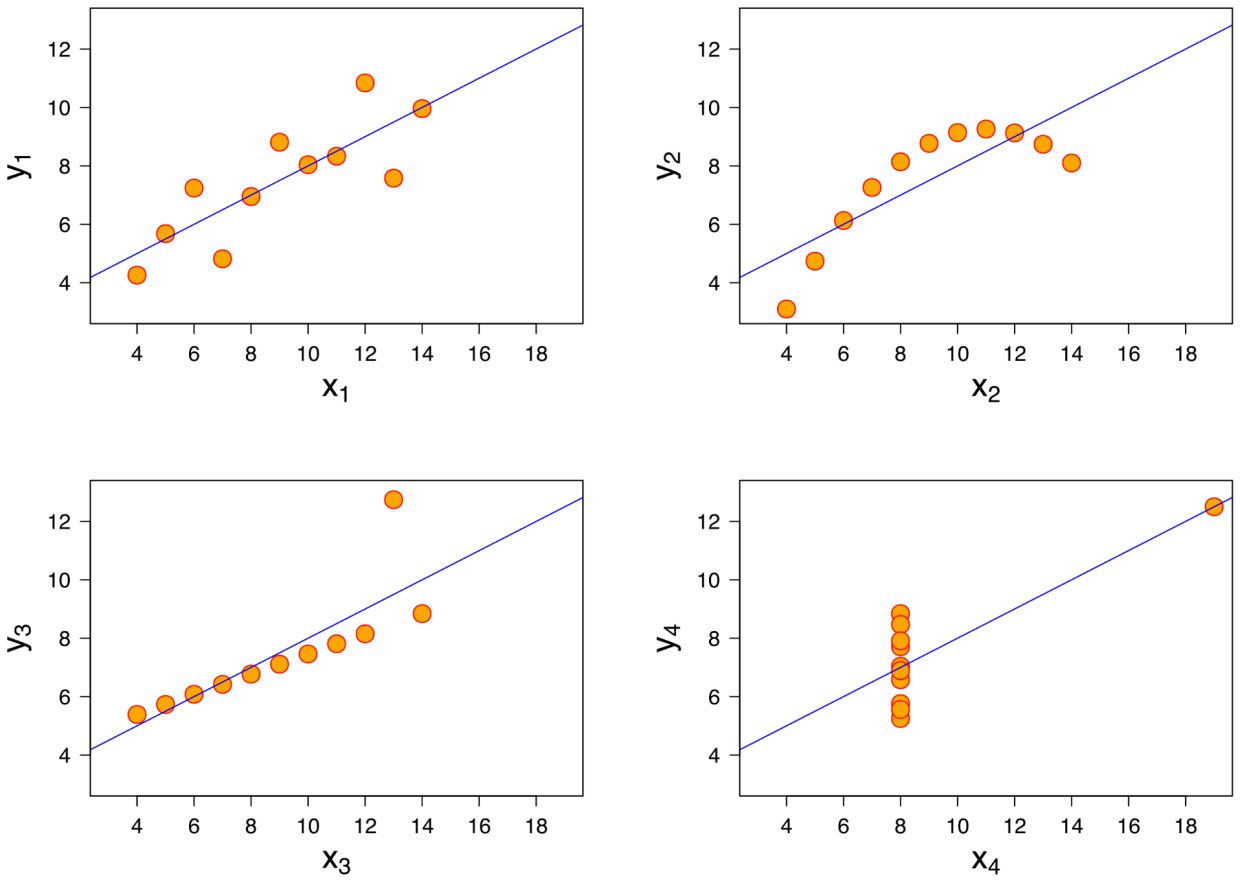
Caption

For all 4 datasets:

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x: s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y: s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression: R^2	0.67	to 2 decimal places

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.

Graphical representations of 4 datasets



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.

————————————————————————————————————

3. What is Pearson's R?

Answer: The Pearson correlation coefficient is a correlation coefficient that measures linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationships or correlations.

Formula for Pearson Coefficient:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Where

Cov is covariance
σ_X is the standard deviation of x
σ_Y is the standard deviation of y

————————————————————————————————————

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

Answer: Scaling is one of the data preprocessing step in machine learning. Algorithms that compute the distance between the features are biased towards numerically larger values if the data is not scaled.

Algorithms are fairly insensitive to the scale of the features. Also, feature scaling helps machine learning, and deep learning algorithms train and converge faster.
Normalization or Min-Max Scaling is used to transform features to be on a similar scale.

It is calculated as:

X_new = (X - X_min)/(X_max - X_min)

This scales the range to [0, 1] or sometimes [-1, 1].

Some features of **Normalization**:

1) Minimum and maximum value of features are used for scaling
2) It is used when features are of different scales.
3) Scales values between [0, 1] or [-1, 1]
4) It is really affected by outliers

Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.

It is calculated as:

X_new = (X - mean)/Std

Some features of **Standardization**:

1)Mean and standard deviation is used for scaling
2)It is used when we want to ensure zero mean and unit standard deviation
3) It is not bounded to a certain range
4)It is much less affected by outliers

———————————————————————————————————————

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

Answer: VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity.
It is calculated as:
VIF =1/(1-(R^2))
If there is a perfect correlation between variables, then r will be equal to 1 which will imply that VIF = Infinity.

———————————————————————————————————————

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

Answer: A Q–Q plot (quantile-quantile plot) is a probability plot for comparing two probability distributions by plotting their quantiles against

each other.A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). This defines a parametric curve where the parameter is the index of the quantile interval.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the identity line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x.

It 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. They can also be used to compare collections of data, or theoretical distributions.