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?**
   Yr, working day, Saturday and September are some of the categorical variables that have an impact on the dependent variable.

2. **Why is it important to use drop_first=True during dummy variable creation?**
   If we don't use drop_first=True, then n dummy variables will be created. This could result in added overload to the machine when performing calculation incase of a huge dataset. Also, there could be the possibility of increased multicollinearity.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   Observing the pairplot, the highest correlation is observed in the "temp" column. Then followed by atemp column.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   We can perform the following checks**:**
   - Normality of residual.
   - Homoscedasticity
   - No Significant multicollinearity

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

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression tries to come up with a Linear function that describes the relationship between the dependent (X) and independent (Y) variables. The function tries to explain the variance in Y using a linear combination of weighted X variables. As it is a Regression technique, Y is always numerical.

The function takes the form:

$Y=\beta0+\beta1X1+\beta2X2+...+\beta pXp+\epsilon$

Where, β0 is a constant intercept and is the ϵ error term. The function can be interpreted as: the predicted value of Y increases by β1 for a unit increase in X1, given other X values remaining constant.

It generates the above function by trying to fit a line (hyper-plane for higher dimensions) through all the data points and minimizing the prediction error. There are two methods for this optimization:

- Ordinary Least Squares
- Gradient Descent

Linear Regression makes some assumptions about the data and relationship between X and Y:

- There is a linear relationship between X and Y
- Error terms are normally distributed with mean zero
- Error terms are independent of each other
- Error terms have constant variance (homoscedasticity)

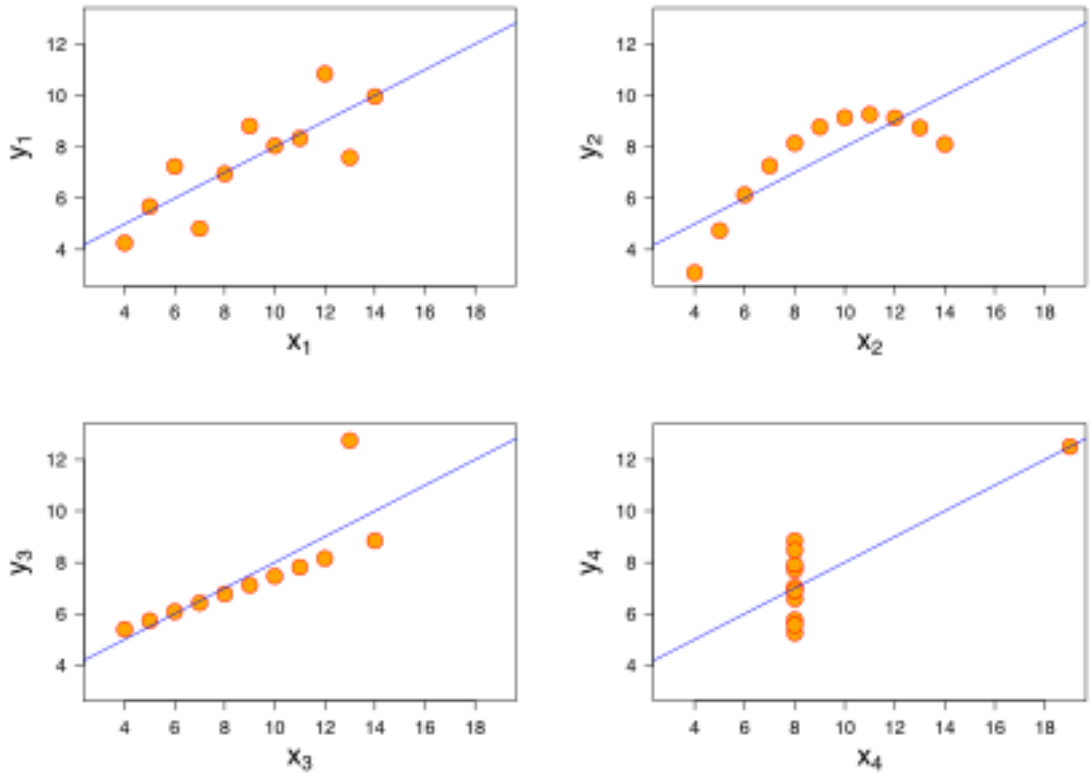
2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet they have very different distributions and appear distinct when graphed. This phenomenon was first presented by the British statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before analyzing it and to illustrate how relying solely on summary statistics can be misleading.

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.

Source:Wikipedia

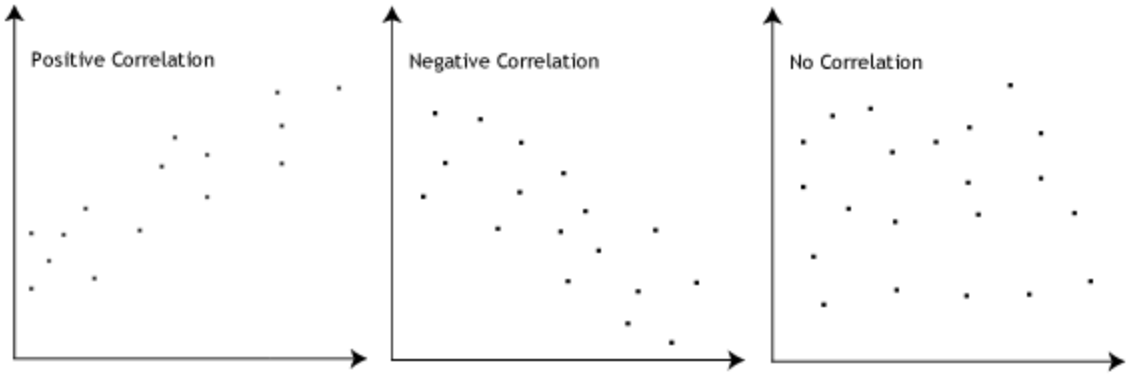
Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x: s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y: s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient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



3. What is Pearson’s R?

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of a linear relationship between two variables. It is represented by the symbol \(r \) and ranges from -1 to 1:

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



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

Scaling is a step in data analysis and machine learning where the values of variables are transformed to fit within a specific range. It is performed to ensure that the variables fit into comparable scales, which can help in easier modeling.

Scaling is Performed for below reasons

- Normalization of Variables
- Improved Convergence
- Distance-based Algorithms
- Regularization
- Visualization

Normalized Scaling vs. Standardized Scaling:
Normalized Scaling also known as Min-Max Scaling is used to transform the data to fall within a specific range, typically between 0 and 1.

Why Normalized Scaling is Used

- To preserve the original distribution of the dataset.
- It's useful when the distribution of the data is unknown.

Standardized Scaling also known as Z-score normalization is used to transform the data to have a mean of 0 and a standard deviation of 1.

Why Standardized Scaling is Used

- It is less sensitive to outliers compared to Min-Max scaling.
- It preserves the original shape of the dataset distribution.

Difference Between Normalized Scaling and Standardized Scaling:

- Normalized scaling (Min-Max) transforms the data within a specific range, while standardized scaling (Z-score) transforms the data to have a mean of 0 and a standard deviation of 1.
- Normalized scaling is useful when preserving the original distribution of the data is important, while standardized scaling is robust to outliers and ensures variables are on a similar scale without binding them to a specific range.

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

VIF = infinity is observed only in case of perfect correlation. Presence of large values of VIF indicates that there is a high correlation between the variables.

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

6. 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 used to determine if two data sets come from populations with a common/similar distribution.

Use of Q-Q plot:

- A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset.
- 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.

Importance of Q-Q plot:

- When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified.
- If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale.
- If two samples do differ, it is also useful to gain some understanding of the differences.
- The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.