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?**

   The categorical variables yr, weathersit, season and month have an impact on the target variable. With the yr variable having the highest impact among them (second highest overall).

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

   Pandas get_dummies function creates one variable per label in the column. But k labels only need k-1 columns to represent all of the information. Hence, we drop the column corresponding to the first label.

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

   Variable atemp has the highest correlation with target variable at 0.64.

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

   After training the model, the following checks were made:
   - Normality of residuals
   - Homoscedasticity of residuals

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

   The top three features are:
   - hum_by_temp
   - yr
   - season_3

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:

   $$Y=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p+\epsilon$$

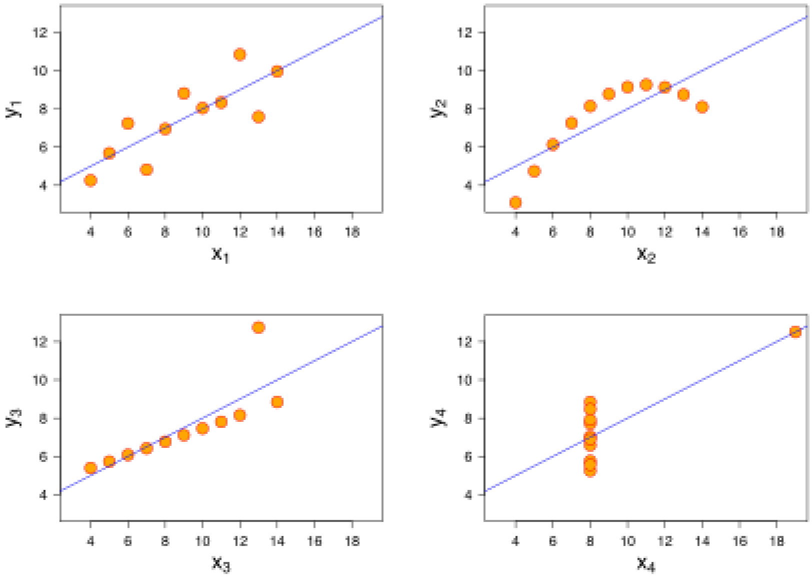
   Where, β_0 is a constant intercept and is the ϵ error term. The function can be interpreted as: the predicted value of Y increases by β_1 for a unit increase in X1, given other X values remaining constant.

   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

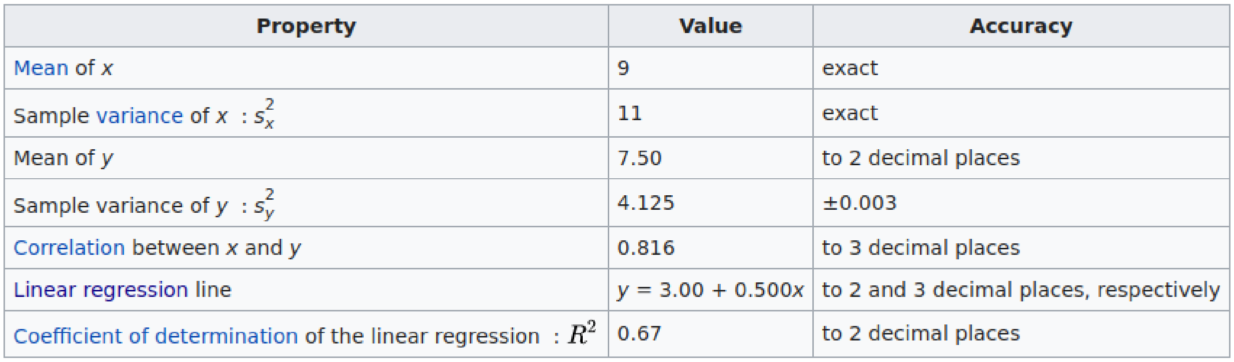
   Linear Regression makes some assumptions about the data and relationship between X and Y:
   - There is a linear relationship between X and Y
   - Error terms are normally distributed with mean zero
   - Error terms are independent of each other
   - Error terms have constant variance (homoscedasticity)

2. **Explain the Anscombe's quartet in detail.**

   Anscombe's quartet 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 before analyzing it and the effect of outliers and other influential observations on statistical properties.



For all four datasets:



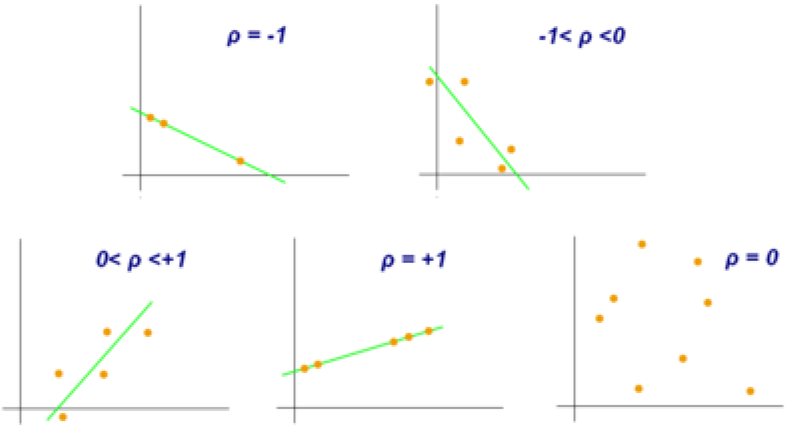
Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x* : s_x^2	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y* : s_y^2	4.125	±0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear regression line	*y* = 3.00 + 0.500*x*	to 2 and 3 decimal places, respectively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

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 statistical properties for describing realistic datasets.

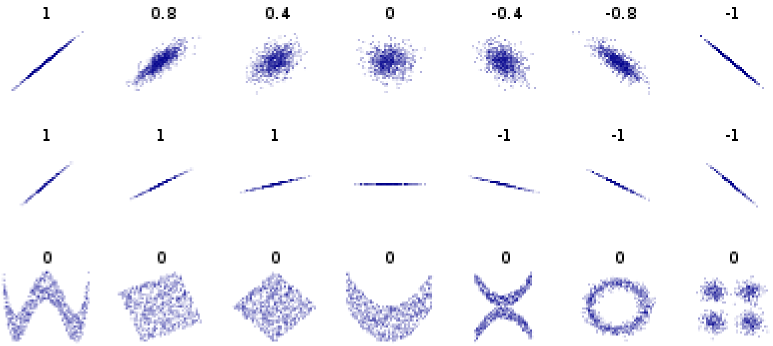
Source: Anscombe's quartet

3. What is Pearson's R?

In statistics, the Pearson correlation coefficient (PCC) — also known as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), the bivariate correlation, or colloquially simply as the correlation coefficient — is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, 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 relationship or correlation.



Examples of scatter diagrams with different values of correlation coefficient (*ρ*)



Several sets of (*x*, *y*) points, with the correlation coefficient of *x* and *y* for each set. Note that the correlation reflects the strength and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of *Y* is zero.

The correlation coefficient ranges from −1 to 1. An absolute value of exactly 1 implies that a linear equation describes the relationship between X and Y perfectly, with all data points lying on a line. The correlation sign is determined by the regression slope: a value of +1 implies that all data points lie on a line for which Y increases as X increases, and vice versa for -1. A value of 0 implies that there is no linear dependency between the variables.

Source: Pearson correlation coefficient

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

   Scaling is the process of bringing the values for all variables within a similar range. The reasons for performing scaling are as follows:

   - It makes model interpretation easier. As all the variables are in the same value range, their coefficients can be compared against each other for variable importance.
   - For optimization algorithms like Gradient Descent, it makes the convergence faster and more stable as the step size is the same for all variables.

   There are two major ways to scale variables:

   - Normalization: Also called Min-Max Scaling. Here, we bring all the values of the variables within 0-1 range, with maximum value being mapped to 1 and minimum mapped to 0. This does not retain the shape/spread of the data points.
   - Standardization: Here, we standardize the values by subtracting mean and dividing by standard deviation. This retains the shape/spread of the data points.

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

   The Variance Inflation Factor (VIF) is given by:

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

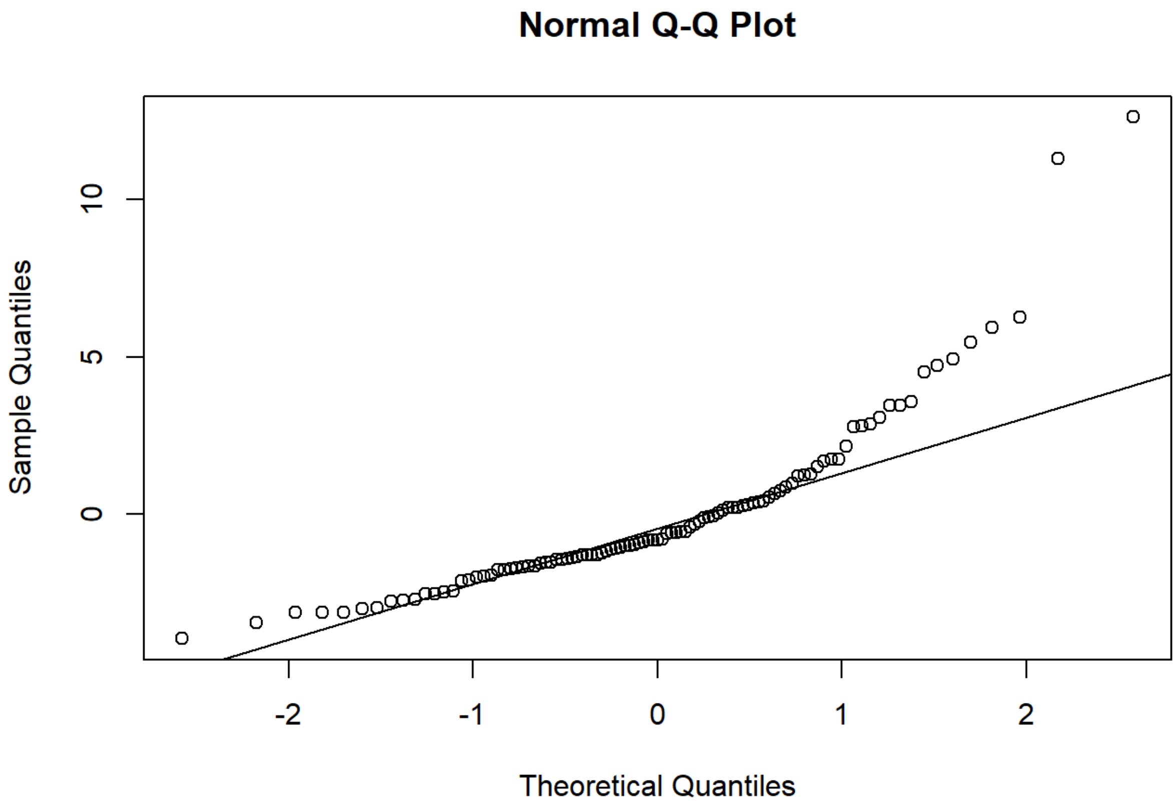
   When the R2 value is 1, the VIF value becomes infinite. This signifies that the variable is being completely explained (100% R2) by one or more variables in the data. Or, the data has perfect multicollinearity.

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

   In statistics, a Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. First, the set of intervals for the quantiles is chosen. 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). Thus the line is a parametric curve with the parameter which is the number of the interval for the quantile.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the 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.

Q-Q plots are used to check the normality of residuals by plotting the residuals against the normal distribution. Q-Q plots take your sample data, sort it in ascending order, and then plot them versus quantiles calculated from a theoretical distribution, in this case the normal distribution. If the points fall pretty closely along the line, the data are normal.



This plot looks pretty good as each point falls pretty close to the 45 degree line but it does show two points pretty far from the line.

Source: Q–Q plot, Normal Q-Q Plot for Linear Regression