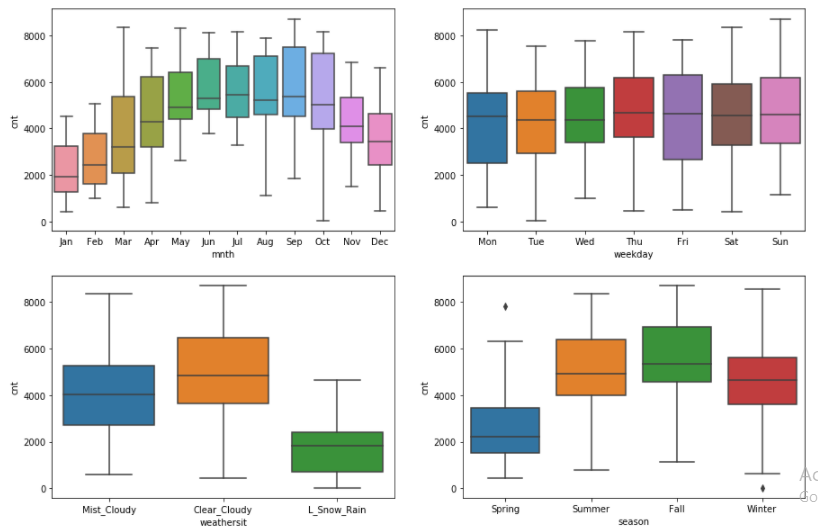
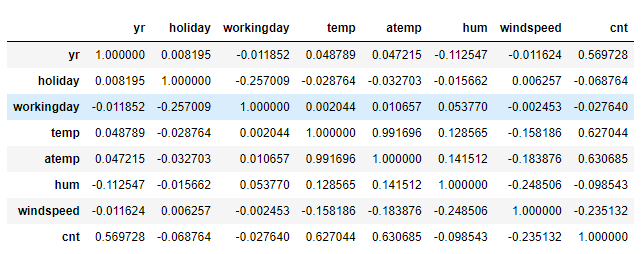
**Assignment-based Subjective Questions**

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**After analysing the categorical variables, we could infer that the Bike rentals & Sales were influenced by the Weather and Season categories. We observed that in the months of June-October, it showed relatively higher sales. Also, ‘Clear’ skies were favoured by the customers availing bike rental services. Fall & Summer were the most preferred seasons for availing bike rentals.



1. **Why is it important to use drop\_first=True during dummy variable creation?**  
   Using drop\_first=True helps with tackling the redundancy of the variables. Consider the following two examples :-  
   **Set 1:**  
    **Column #1 Column #2 Column #3**Flag 1 0 0Flag 0 1 0Flag 0 0 1 **Set 2:  
    Column #1 Column #2**Flag 1 0 Flag 0 1 Flag 0 0   
     
   Both of these two sets infer the same meaning but Set 2 helps us save space on the disk and makes computation relatively faster while model building step.
2. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**Looking at the Pair-Plot among the numerical variables, ‘Atemp’ was found to be having the highest correlation with the target variable ‘Cnt’



1. **How did you validate the assumptions of Linear Regression after building the model on the training set?**  
     
   After building the training set, we validated the assumptions of Linear Regression by checking **Multicollinearity** using Variance Inflation Factor (VIF). We also checked for **Normality of Residuals** and found a normal distribution of errors centred on mean of 0.
2. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**  
     
   Based on the final model, the following are considered to be contributing significantly to the demand of the shared bikes :-  
   **- Season  
   - Weekday  
   - Weathersit**

**General Subjective Questions**

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

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y (output). Hence, the name is Linear Regression. In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

**Hypothesis function for Linear Regression:**



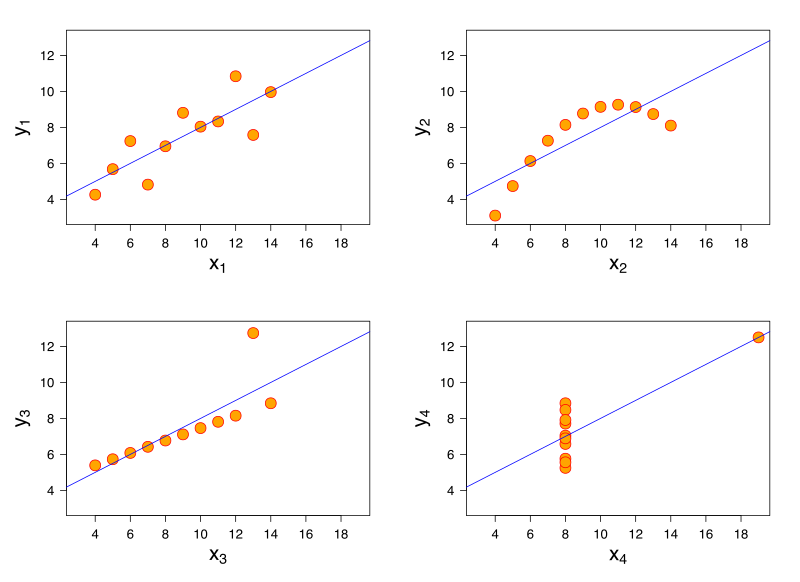
While training the model we are given:

**x:** input training data (Univariate – one input variable(parameter))

**y:** labels to data (supervised learning)

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ1 and θ2 values.

**θ1:** intercept  
**θ2:** coefficient of x.

**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](https://en.wikipedia.org/wiki/Statistician) [Francis Anscombe](https://en.wikipedia.org/wiki/Francis_Anscombe) to demonstrate both the importance of graphing data when analysing it, and the effect of [outliers](https://en.wikipedia.org/wiki/Outlier) and other [influential observations](https://en.wikipedia.org/wiki/Influential_observations) on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."  
  
  
For all four datasets:

• 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) is not distributed normally; 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 distribution 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.

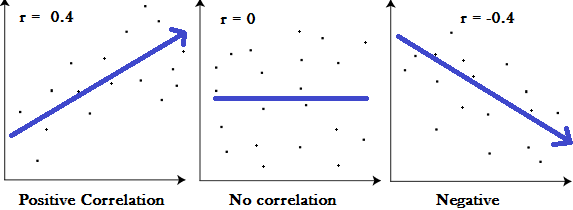
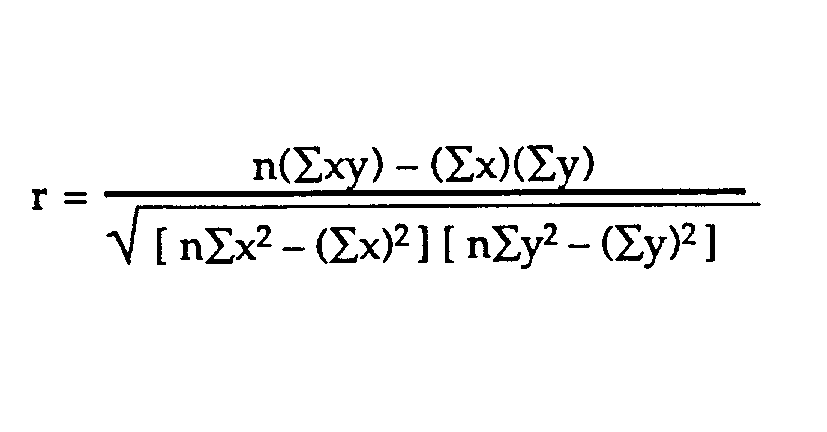
**3. What is Pearson’s R?**

**Correlation coefficients** are used in statistics to measure how strong a relationship is between two variables. There are several types of correlation coefficient, but the most popular is Pearson’s. **Pearson’s correlation** (also called Pearson’s *R*) is a **correlation coefficient** commonly used in linear regression.

Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

• 1 indicates a strong positive relationship.

• -1 indicates a strong negative relationship.

• A result of zero indicates no relationship at all.   
  
  
 *(Graphs showing a correlation of -1 (a negative correlation), 0 and +1 (a positive correlation))*The absolute value of the correlation coefficient gives us the relationship strength. The larger the number, the stronger the relationship. For example, |-.75| = .75, which has a stronger relationship than .65.  
  
One of the most commonly used formulas in stats is Pearson’s correlation coefficient formula. If you’re taking a basic stats class, this is the one you’ll probably use:  
  


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

**Feature Scaling** is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step.

By scaling your variables, you can help compare different variables on equal footing.

Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, many classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.  
  
Another reason why feature scaling is applied is that gradient descent converges much faster with feature scaling than without it.

Normalization V/s Standardization

**Normalization** is good to use when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.

**Standardization**, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).   
Since VIF = 1/Tolerance, it indicates 0 tolerance which means 0 collinearity, i.e because Tolerance is a measure of Collinearity.

The Variance Inflation Factor (VIF) measures the impact of collinearity among the variables in a regression model. The Variance Inflation Factor (VIF) is 1/Tolerance, it is always greater than or equal to 1. There is no formal VIF value for determining presence of multicollinearity. Values of VIF that exceed 10 are often regarded as indicating multicollinearity, but in weaker models values above 2.5 may be a cause for concern. In many statistics programs, the results are shown both as an individual R2 value (distinct from the overall R2 of the model) and a Variance Inflation Factor (VIF). When those R2 and VIF values are high for any of the variables in your model, multicollinearity is probably an issue. When VIF is high there is high multicollinearity and instability of the b and beta coefficients.

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

The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. It’s just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.  
  
A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that’s roughly straight. Here’s an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.

