# 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?**

**Ans:** To understand the effect of categorical variables on the dependent variable, I trained two separate models on (i) only numerical variables (ii) numerical + categorical variables. The essential summary statistics of the two models is shown below.

| **Model train on only numerical variables** | | **Model trained on numerical+categorical variables** | |
|---|---|---|---|
| R-squared: | 0.711 | R-squared: | 0.810 |
| Adj. R-squared: | 0.707 | Adj. R-squared: | 0.798 |
| Prob (F-statistic): | 5.17e-132 | Prob (F-statistic): | 7.62e-153 |

We can clearly see that the model's performance has improved with the addition of categorical variables. The increase in Adj. R-squared suggests that the added categorical features are important. Furthermore, many categorical variables are also present in the final model selected in this project (ie. Model_6 in jupyter notebook). The p-value of these categorical variables is less than 0.05, indicating statistical significance of the categorical variables.

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

**Ans:** drop_first=True removes the first column created in dummy variables. If a categorical variable has m levels, then only m-1 dummy variables are sufficient to represent all the levels. Information contained in m-1 dummy variables is the same as the information in m variables, making one column redundant. Hence, we are free to drop any column it doesn't matter which one we drop. For convenience we choose to drop first. The redundant column is highly correlated with others; hence dropping it reduces the correlations among dummy variables. It also reduces the number of variables in the models. This is particularly important when we have a large number of categorical variables.

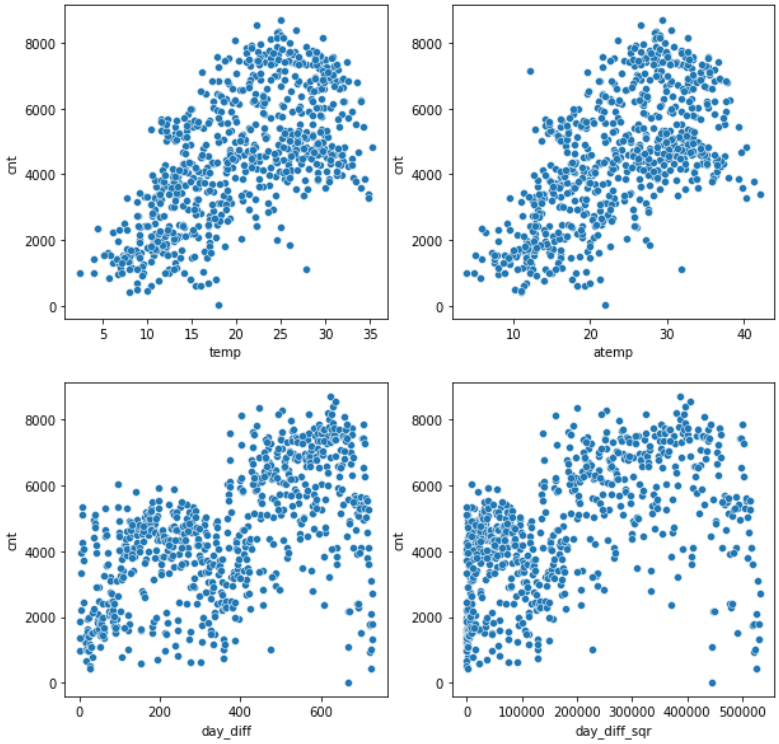
**3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

**Ans:** "registered" variable has the highest correlation with the target variable.

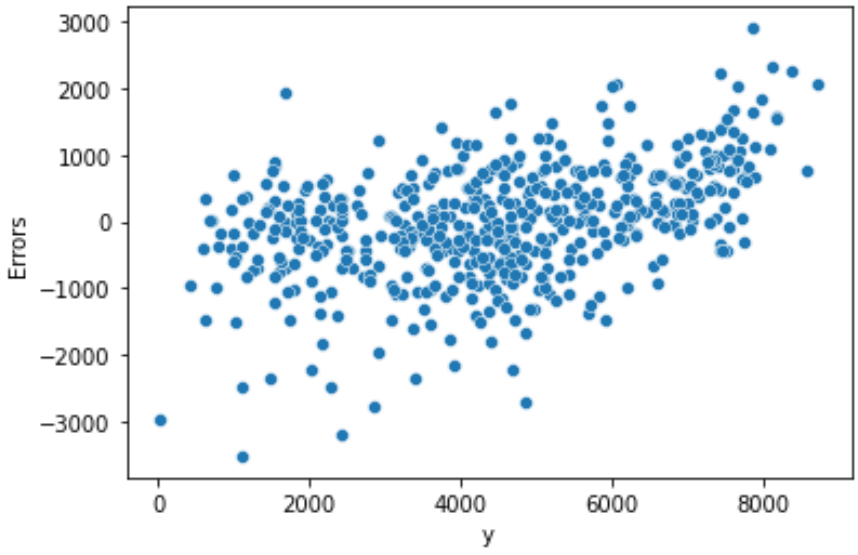
**4. How did you validate the assumptions of Linear Regression after building the model on the training set?**

**Ans:** I validated the following assumptions of Linear Regression by creating appropriate plots as shown below.

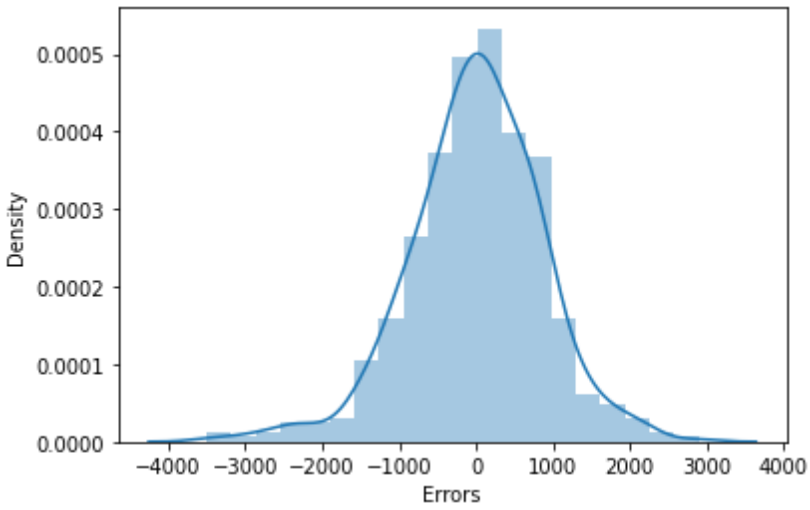
- **There exists linear relationship between X and Y:** Variables "temp" and "atemp" show approximately linear relationship with the target variable. Whereas variables "day_diff" and "day_diff_sqr" show a weak linear trend with the target variable as shown in the plots below. Hence, we can say that this assumption is not completely satisfied.



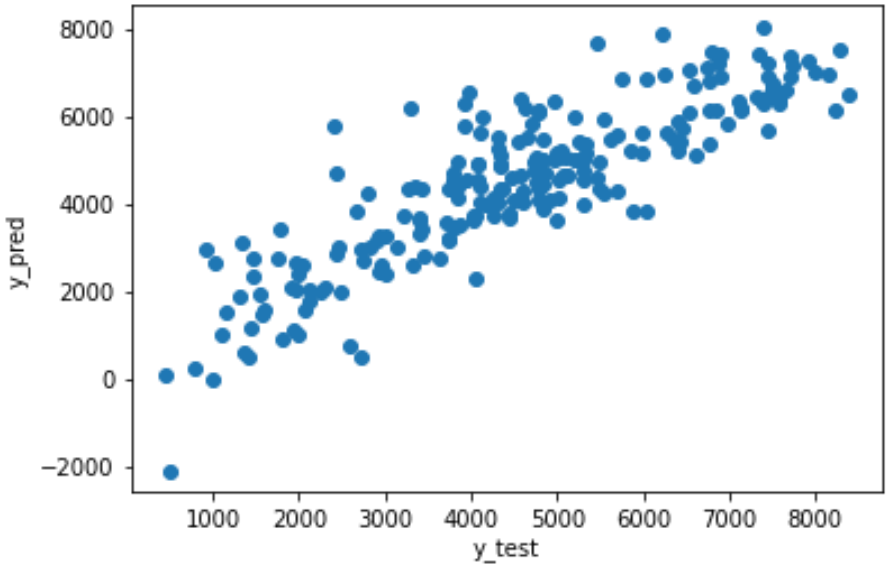

- **Error terms are independent of each othe:** I created a plot between residual error and target as shown below to verify this assumption. We see no visible relationship between error and target variable validating assumption of independence.




- **Error terms are normally distributed:** This assumption was verified through the residual analysis using the predictions from the model. Plot below shows the approximately normal distribution of errors, validating the assumption.




- **Error terms have constant variance (homoscedasticity):** To validate this assumption, I evaluated the model on the test set and created a plot between the predicted and true value of the target variable shown below. From the plot, it can be seen that the variance of errors is approximately constant.




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

**Ans:** Based on the absolute value of the coefficients, these variables contribute significantly towards the demand of the shared bike: "temp", "day_diff", "weathersit_3"

- 'temp' is positively correlated with the bike demand, suggesting bike demand is higher during summer.
- 'day_diff' shows that bike demand has an increasing trend with the time so company should increase the number of bikes every year.
- 'weathersit_3' has a negative coefficient indicating a decrease in bike demand during rainy and snowy weather, its because people tend to avoid travelling in rainy and snowy conditions.

# General Subjective Questions

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

**Ans:** Regression involves modelling target variable based on independent variables. Regression is used to understand the relationship between target and independent variables. Different regression techniques have been developed based on the number of independent variables and the type of relationship between independent and dependent variables. Unlike classification, the dependent variable is always continuous in regression analysis. Linear regression is a regression technique (model) that assumes a linear relationship between predictor and the target variable. It is one of the simplest machine learning models used when the target variable depends linearly on independent variables. The target variable is represented as a linear combination of independent variables in linear regression. Based on the number of independent variables, there are two types of linear regression (i) Simple Linear Regression (ii) Multiple linear regression

- **Simple Linear Regression:** The linear regression model with only one independent variable is known as a simple linear regression. The relationship between the target and independent variable is given by the following equation of a straight line

$$y = \beta_0 + \beta_1 x$$

- **Multiple Linear Regression:** When there are several independent variables in the linear regression model, it is called multiple linear regression.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Linear regression involves four assumptions:

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

The goal of linear regression is to find the best-fit line having the minimum error of prediction. Linear regression algorithm involves following steps.

- **Data Understanding and preparation:** This step is not strictly part of the algorithm. However, it's a crucial step in modelling. This step addresses the data quality issues such as missing values, null values, incorrect datatypes, etc. New variables are created based on domain knowledge. Then, categorical variables are dealt with using one-hot encoding, label encoding and dummy variables. Data is then split into a training set and test set. Next, continuous variables are scaled appropriately. We have training data ready for modelling at the end of this step.

- **Training:** Training involves defining a cost function. The cost function in linear regression is Mean Squared Error (MSE).

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_{true} - y_{pred}\right)^2$$

Where,
N is the number of training points.
$y_{true}$ is the true values of the target variable.
$y_{pred} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$

To identify the best-fit line, we minimize the MSE using gradient decent. The output of gradient descent is the set of coefficients (betas) giving the lowest MSE. Gradient decent is the process of optimizing the values of coefficients by iteratively minimizing the MSE on training data. It starts with the random values for each coefficient. The MSE is calculated on the training data. The coefficients are updated in the direction of minima using learning rate as scaling factor. The process is repeated until MSE below some threshold. Gradient decent works very well on the large dataset. However, the closed-form solution exists for linear regression as given below.

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- **Variable selection:** In this step, redundant and useless variables are discarded through hypothesis testing and monitoring VIF, $R^2$, adjusted $R^2$. Less number features result in an interpretable and simple model.

- **Evaluation on the Test set:** This is the final step in which we evaluate the best model on the test set to understand it generalization power. Generally, MSE and $R^2$ are used as evaluation metrics. Before evaluation on the test set, it is first scaled appropriately.

**2. Explain the Anscombe's quartet in detail.**
**Ans:** Anscombe's quartet was constructed by a statistician Francis Anscombe in 1973 to illustrate the importance of visualization before analyzing and model building, the effect of outlier and other observations on statistical properties. Anscombe's quartet constitutes four data sets with nearly identical summary statistics. However, they have very different distributions and looks completely different when plotted. He created four datasets (Table 1) to demonstrate the dangers of summary statistics. Each dataset consists of eleven (x,y) pairs as follows:

*Table 1*

| **I** | | **II** | | **III** | | **IV** | |
|---|---|---|---|---|---|---|---|
| *x* | *y* | *x* | *y* | *x* | *y* | *x* | *y* |
| 10 | 8.04 | 10 | 9.14 | 10 | 7.46 | 8 | 6.58 |
| 8 | 6.95 | 8 | 8.14 | 8 | 6.77 | 8 | 5.76 |
| 13 | 7.58 | 13 | 8.74 | 13 | 12.74 | 8 | 7.71 |
| 9 | 8.81 | 9 | 8.77 | 9 | 7.11 | 8 | 8.84 |
| 11 | 8.33 | 11 | 9.26 | 11 | 7.81 | 8 | 8.47 |
| 14 | 9.96 | 14 | 8.1 | 14 | 8.84 | 8 | 7.04 |
| 6 | 7.24 | 6 | 6.13 | 6 | 6.08 | 8 | 5.25 |
| 4 | 4.26 | 4 | 3.1 | 4 | 5.39 | 19 | 12.5 |
| 12 | 10.84 | 12 | 9.13 | 12 | 8.15 | 8 | 5.56 |
| 7 | 4.82 | 7 | 7.26 | 7 | 6.42 | 8 | 7.91 |
| 5 | 5.68 | 5 | 4.74 | 5 | 5.73 | 8 | 6.89 |

All the summary statistics of these datasets are nearly identical:

- The mean of x and y for each dataset is 9 and 7.50 for each dataset.
- The variance of x and y for each dataset is 11 and 4.12 for each dataset.
- The correlation between x and y is 0.816 for each dataset.
- Linear regression for each dataset follows the equation: $y = 0.5x + 3$.

Hence, these datasets appear almost identical, but we see a very different picture when we plot them on x-y plane (Figure 1).

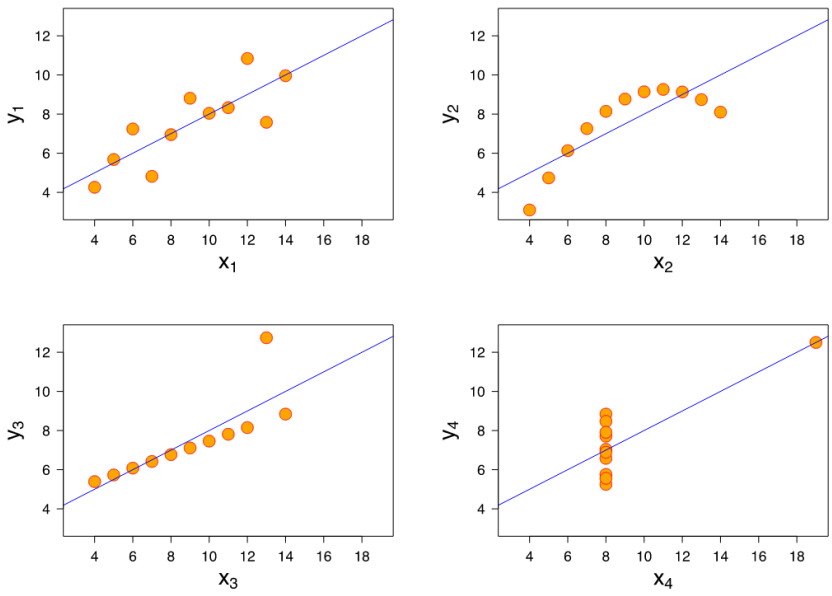



*Figure 1*

We can see that in the dataset I, y follows a roughly linear relationship with x. In dataset II, y has non-linear relationship with x. In dataset III, y seems to follow a tight linear relationship with x except for one outlier. In dataset IV, x is constant again with one outlier. This tells us the importance of visualizing data before modelling. Plotting data gives us a clear picture of the distribution and relationship and can also help indemnify various anomalies present in the data.

**3. What is Pearson's R?**

**Ans:** Pearson's R also known as Pearson correlation coefficient (PCC), the Pearson product-moment correlation coefficient (PPMCC), the bivariate correlation or simply as the correlation coefficient. Pearson's R is a measures the degree of linear correlation between two sets of data. It is defined as the ratio between covariance of two variable and the product of their standard deviations. The concept of covariance comes from probability theory; it measures the joint variability of two random variables. The formulas of Pearson's R for the population and sample is given below.

For population: $r = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$

Where,
$cov$ is the covariance

$\sigma_X$ is the standard deviation of X
$\sigma_Y$ is the standard deviation of Y

For sample: $r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$

Where,

$n$ is the sample size

$x_i$, $y_i$ are the individual sample points

Pearson's R is can be viewed as the normalized measure of the covariance. Hence, it lies between -1 and 1. Pearson's R measures the strength of association and direction between two linear variables. Simply, Pearson's R defines the effect of change in one variable when the other variable is changed. When the data points fall very close to the best-fit line, Pearson's R tends to have high values (-1 or 1), suggesting a strong association between two linear variables. The further the data points move away from the best-fit line, the weaker the strength of the linear relationship. The direction of the best-fit line determines the sign of Pearson's R. When the values of Pearson's R is positive, the best-fit line also has a positive slope, indicating a positive relationship between variables. This means an increase in the value of one variable will lead to an increase in the value of the other variable. A negative value of Pearson's R results in negative slope of best-fit line, indicating negative relationship. This means an increase in the amount of one variable leads to a decrease in the value of another variable. These concepts are shown in Figure 2.

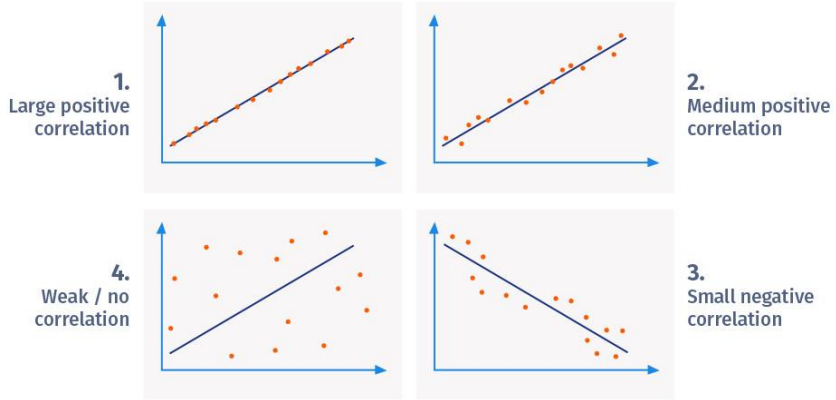



*Figure 2*

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

**Ans:** The scaling or feature scaling is the transformation of the values of independent variables such that the resulting dataset has specific statistical properties. Some machine learning algorithms are sensitive to the range of features. If there is a vast difference in the range, then the algorithm may give more preference to the higher ranging features and start dominating the predictions. For example, in a linear regression, let's say some features range in thousands and others in tens. Then, the model would reduce the magnitude coefficients for the features having values in thousands and increase for the features having values in tens. This makes the model less interpretable as we can't compare the coefficient to understand the features' relative importance. Another reason feature scaling is required is that machine learning algorithms (eg. linear regression, logistic regression, neural network) that use gradient descent as an optimization technique converge much faster with feature scaling than without it.

There are two popular methods of scaling:

| **Normalization** | **Standardization** |
|---|---|
| • It is also known as min-max scaling. Normalization is a scaling technique in which values are shifted and rescaled such that they end up ranging between [0,1] or [-1,1]. It is also possible to normalize over different intervals depending on business needs. The general formula to rescale between arbitrary values [a,b] is shown below.<br><br>$x' = a + \frac{(x-\min(x))(b-a)}{\max(x)-\min(x)}$<br><br>• Normalization shifts and compresses values typically in [0,1] or [-1,1] interval.<br>• Normalization is good to use when the distribution of data doesn't follow a Gaussian distribution<br>• The impact of outliers is very high in the normalization | • Standardization transforms the data to have zero mean and a variance of 1. The formula for standardization is shown below.<br><br>$x' = \frac{x-\bar{x}}{\sigma}$<br><br>• Standardization scales values based on the variance<br>• In standardization, upper bound and lower bound of the data are not predefined.<br>• Standardization can be helpful in cases where the data follows a Gaussian distribution.<br>• Outlier as not affected by standardization. |

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

**Ans:** VIF becomes infinite when the independent variable is completely described as the linear combination of other independent variables. When all the variance in an independent variable is fully explained by other variables, $R^2$ becomes one and according to the VIF formula shown below, it becomes infinite.

$$VIF = \frac{1}{1 - R^2}$$

When $R^2$ is one the denominator becomes 0 making VIF infinite.

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

**Ans:** The Q-Q plot is a quantile-quantile plot. Q-Q plot is a graphical method of identifying whether two data sets come from populations with the same distribution. Q-Q plot is a plot of the quantile of the first data set against the quantile of the second data set.

Advantages:

- It is used to check if two data sets come from a population with a common distribution.
- It can be used to verify if the data comes from some theoretical distribution such as normal, exponential or uniform distribution.
- It can be used to detect if the two data sets come from the populations whose distributions differ only by a shift in location
- It can be used with data set having different sample sizes.

- It can be used to understand the distributional aspects such as shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.

In linear regression, the Q-Q plot could be used to verify whether errors are normally distributed or not. It can also be used to check if the training and test set comes from same the distribution or not.