# 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?**
**Based on the boxplot for categorical variables with 'cnt' variable:**

- The working day and holiday box plots indicate that more bikes are rent during normal working days than on weekends or holidays.
- The weathersit box plots indicates that more bikes are rent during Clear, Few clouds, Partly cloudy weather.
- The season box plots indicates that more bikes are rent during fall season.
- The month box plots indicates that more bikes are rent during september month.
- The year box plots indicates that more bikes are rent during 2019.
- The weekday box plots indicates that more bikes are rent during saturday.

**2. Why is it important to use drop_first=True during dummy variable creation?**
As there is no need to define all different levels. If you drop a level, you will still be able to explain all the levels so we
use drop_first=True during dummy variable creation.

**3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
We have observed from pair plot that 'temp' 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?**

- I checked the error terms are also normally distributed with mean zero (which is infact, one of the major assumptions of linear regression), for this we will plot the histogram of the error terms.
- I also find that there is no clear pattern and distribution of error terms. Also error terms have constant variance which is known as homoscedasticity.
- There is a Linear relationship between independent variable 'temp' and dependent variable 'cnt' and it can be by a scatter and line plot.

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

- **Temp:** As the coefficient of 'temp' is highest and since it is positive so change in Temp variable will directly affect the demand of bikes.
- **weathersit_3:** As the coefficient of 'weathersit_3' is second highest and since it is negative so change in weathersit_3 variable will inversely affect the demand of bikes.
- **Yr:** As the coefficient of 'Yr' is third highest and since it is positive so change in Yr variable will directly affect the demand of bikes.

# General Subjective Questions

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

Linear regression model is defined by the equation y=mx+c.

where y is the dependent variable, x is the independent variable, m is the slope of the line and c is y-intercept.

The Algorithm for linear regression is as follows.

Step 1:

Import the data set and required libraries

```
import pandas as pd
day = pd.read_csv("day.csv")
```

Step 2:

Understanding the data frame.

```
# To display the first 5 day.head()
# To display the last 5 day.tail()
```

Step 3:

Preparing X and y

Putting feature variable to X

Putting response variable to y

Step 4:

Splitting data into train and test

```
from sklearn.cross_validation import train_test_split
df_train,df_test=train_test_split(day, train_size = 0.7, random_state = 100)
```

Step 4:

Performing Linear Regression

```
# import LinearRegression from sklearn
from sklearn.linear_model import LinearRegression
# Representing LinearRegression as lm(Creating LinearRegression Object)
lm = LinearRegression()
#You don't need to specify an object to save the result because 'lr' will take the results of the fitted model.
lm.fit(X_train, y_train)
```

Step 5:

Using SKlearn linear regression model, Visualize the graph (y_predicted vs x(temp))

Step 5:

Coefficients calculation

```
print(lm.intercept_)
print(lm.coef_)
```

Step 6:
Making predictions

```
# Making predictions on the testing Set
 y_pred = lm.predict(X_test)
```

Step 7:
Model evaluation (Plot Actual vs Predicted)

Step 8:
Model evaluation (Plot Error terms)

Step9:
Checking mean square error and R square
Validating Simple Linear Regression Model/ Performance of Regression model by below methods.

a) **coefficient of determination:** denoted R^2 or r^2 and pronounced "R squared", is the proportion of the variance in the dependent variable that is predictable from the independent variable(s).
   R 2 = 1 - (RSS / TSS)
   higher the r2 score better the model

   ```
   from sklearn.metrics import r2_score
   r2_score(y_true= y_test ,y_pred=y_test_pred)
   ```

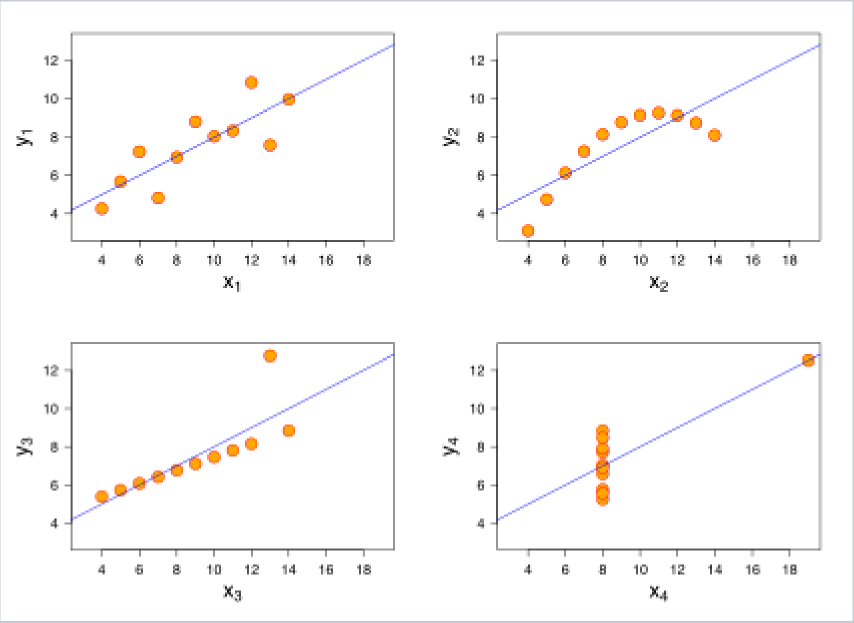
b) **Root Mean Squared Error (RMSE):** Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit.
   lesser the RMSE better the model

   **c) RSS (Residual Sum of Squares): I**n statistics, it is defined as the total sum of error across the whole sample. It is the measure of the difference between the expected and the actual output. A small RSS indicates a tight fit of the model to the data.

   **d) TSS (Total sum of squares):** It is the sum of errors of the data points from mean of response variable.

**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. It is used to demonstrate both the importance of graphing data before analysing it and the effect of outliers and other influential observations on statistical properties.



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

- 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. 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.

- 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?**

Pearson's Correlation Coefficient

Correlation is a technique for investigating the relationship between two quantitative, continuous variables, for example, age and blood pressure. Pearson's correlation coefficient (r) is a measure of the strength of the association between the two variables.

The symbol for Pearson's correlation is "ρ" when it is measured in the population and "r" when it is measured in a sample. Because we will be dealing almost exclusively with samples, we will use r to represent Pearson's correlation unless otherwise noted.

Pearson's r can range from -1 to 1. An r of -1 indicates a perfect negative linear relationship between variables, an r of 0 indicates no linear relationship between variables, and an r of 1 indicates a perfect positive linear relationship between variables.

The first step in studying the relationship between two continuous variables is to draw a scatter plot of the variables to check for linearity. The correlation coefficient should not be calculated if the relationship is not linear. For correlation only purposes, it does not really matter on which axis the variables are plotted. However, conventionally, the independent (or explanatory) variable is plotted on the x-axis (horizontally) and the dependent (or response) variable is plotted on the y-axis (vertically).

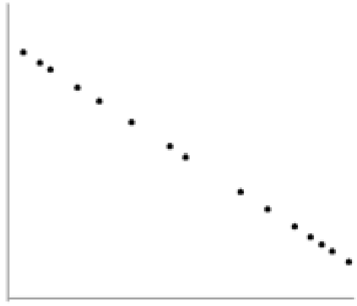
The nearer the scatter of points is to a straight line, the higher the strength of association between the variables. Also, it does not matter what measurement units are used.

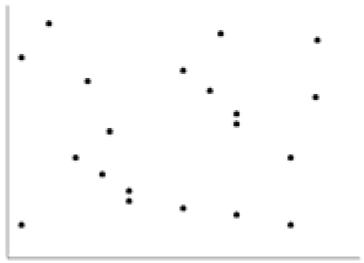
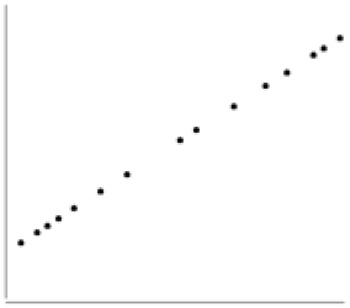
Values of Pearson's correlation coefficient

Pearson's correlation coefficient (r) for continuous (interval level) data ranges from -1 to +1:

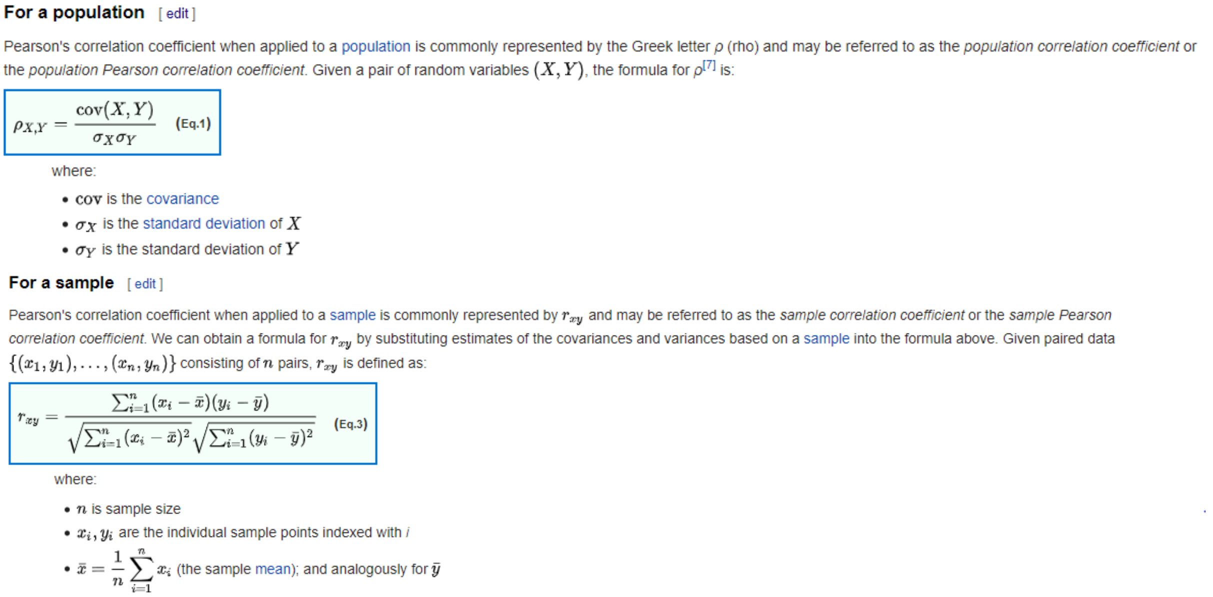
| | |
|---|---|
| r = -1 | data lie on a perfect straight line with a negative slope |
| r = 0 | no linear relationship between the variables |
| r = +1 | data lie on a perfect straight line with a positive slope |

Positive correlation indicates that both variables increase or decrease together, whereas negative correlation indicates that as one variable increases, so the other decreases, and vice versa.

| | | |
|---|---|---|
| r = -1 |  | data lie on a perfect straight line with a negative slope |

| | | |
|---|---|---|
| r = 0 |  | no linear relationship between the variables |
| r = +1 |  | data lie on a perfect straight line with a positive slope |

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.



**For a population** [ edit ]

Pearson's correlation coefficient when applied to a population is commonly represented by the Greek letter $\rho$ (rho) and may be referred to as the *population correlation coefficient* or the *population Pearson correlation coefficient*. Given a pair of random variables $(X, Y)$, the formula for $\rho$[7] is:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \qquad \text{(Eq.1)}$$

where:

- $\operatorname{cov}$ is the covariance
- $\sigma_X$ is the standard deviation of $X$
- $\sigma_Y$ is the standard deviation of $Y$

**For a sample** [ edit ]

Pearson's correlation coefficient when applied to a sample is commonly represented by $r_{xy}$ and may be referred to as the *sample correlation coefficient* or the *sample Pearson correlation coefficient*. We can obtain a formula for $r_{xy}$ by substituting estimates of the covariances and variances based on a sample into the formula above. Given paired data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ consisting of $n$ pairs, $r_{xy}$ is defined as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad \text{(Eq.3)}$$

where:

- $n$ is sample size
- $x_i, y_i$ are the individual sample points indexed with *i*
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

Yes, the following guidelines have been proposed:

| | Coefficient, *r* | |
|---|---|---|
| Strength of Association | Positive | Negative |
| Small | .1 to .3 | -0.1 to -0.3 |
| Medium | .3 to .5 | -0.3 to -0.5 |
| Large | .5 to 1.0 | -0.5 to -1.0 |

Remember that these values are guidelines and whether an association is strong or not will also depend on what you are measuring.

**4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

Most of the times, your dataset will contain features highly varying in magnitudes, units and range. But since, most of the machine learning algorithms use Eucledian distance between two data points in their computations, this is a problem.
If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values that's why scaling is performed.
**Example:** If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

Scaling just affects the coefficients and none of the other parameters, such as t-statistic, F-statistic, p-values and R-squared.

Two major methods are employed to scale the variables: standardisation and MinMax scaling.
**Standardisation:** It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

**MinMax scaling:** This technique re-scales a feature or observation value with distribution value between 0 and 1..
The formulae used in the background for each of these methods are as given below:

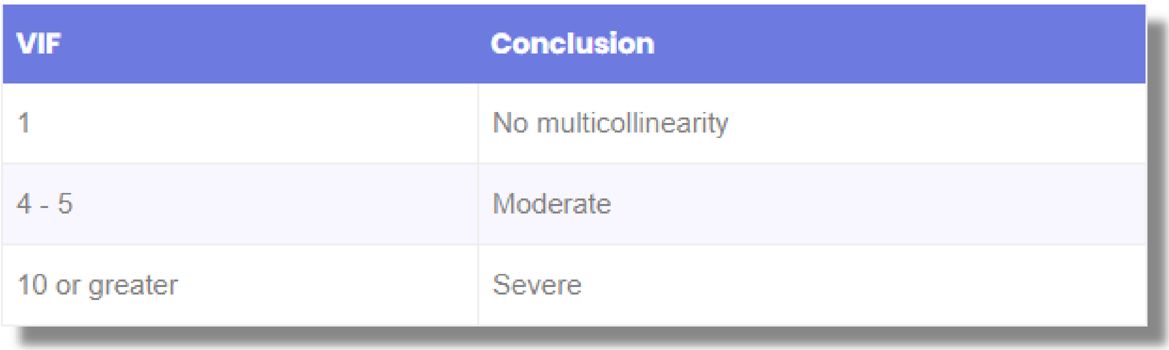
$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

**Difference between Normalization and Standardization:**

- Normalization usually means to scale a variable to have a values between 0 and 1, while standardization transforms data to have a mean of zero and a standard deviation of 1.
- 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 Neighbhours 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? (3 marks)**

Multicollinearity refers to the problem when the independent variables are collinear. Collinearity refers to a linear relationship between two explanatory variables. Two variables are perfectly collinear if there is an exact relationship between the two variables. If the independent variables are perfectly collinear, then our model becomes singular and it would not be possible to uniquely identify the model coefficients mathematically.
One way to address this issue is to check the correlation coefficient between the independent variables and if the correlation coefficient is high (either close to +1 or -1) then we conclude that the variables may be collinear,
VIF ( Variance Inflation Factor) is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.



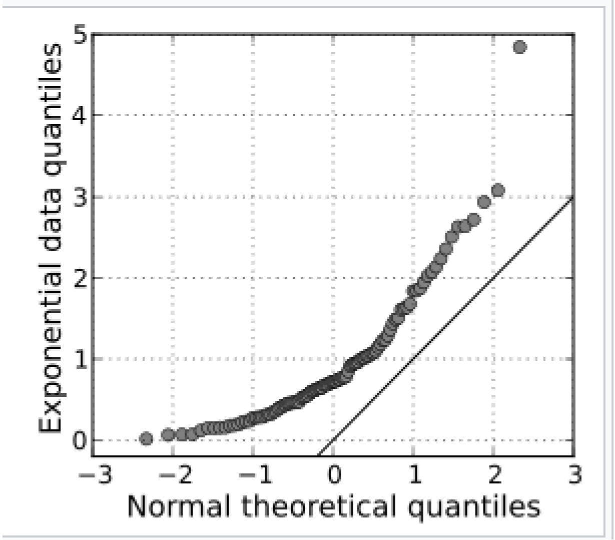
| VIF | Conclusion |
|---|---|
| 1 | No multicollinearity |
| 4 - 5 | Moderate |
| 10 or greater | Severe |

VIF Formula:

**VIF = 1 / (1 – R2)**

In VIF, each feature is regression against all other features. If R2 is more which means this feature is correlated with other features.
When R2 reaches 1, VIF reaches infinity
Once we identify high VIF for features we need to reduce it and we can do it by eliminating some features based on combination of p values and VIF.

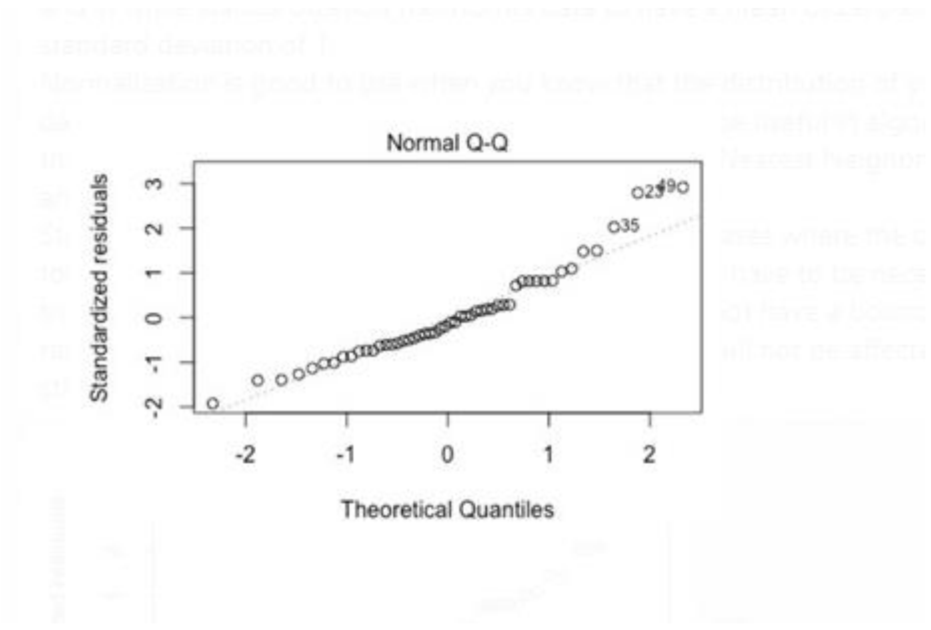
**6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential.
Q Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.



The image above shows quantiles from a theoretical normal distribution on the horizontal axis. It's being compared to a set of data on the y-axis. This particular type of Q Q plot is called a normal quantile-quantile (QQ) plot. The points are not clustered on the 45-degree line, and in fact follow a curve, suggesting that the sample data is not normally distributed.

For a linear regression model, we check if the points lie approximately on the line, and if they don't, the residuals aren't Gaussian and thus the errors aren't either. This implies that for small sample sizes, you can't assume estimator as Gaussian either, so the standard confidence intervals and significance tests are invalid. However, it's worth trying to understand how the plot is created in order to characterize observed violations.

On fitting OLS on a dataset and then analysing the resulting QQ plots.



The q-q plot is formed by:

Vertical axis: Estimated quantiles from data set 1
Horizontal axis: Estimated quantiles from data set 2
Both axes are in units of their respective data sets. That is, the actual quantile level is not plotted. For a given point on the q-q plot, we know that the quantile level is the same for both points, but not what that quantile level actually is.
If the data sets have the same size, the q-q plot is essentially a plot of sorted data set 1 against sorted data set 2. If the data sets are not of equal size, the quantiles are usually picked to correspond to the sorted values from the smaller data set and then the quantiles for the larger data set are interpolated.
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.