# Assignment-based Subjective Questions

**Q1** From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

**Ans -**

**Season**

- Season Fall has maximum Rides while spring has lowest rides

**Weathersit**

Clear weather has max count of rides(1 – code) while heavy rain has no rides(4 – code)

**Weekday**

- There are some data quality issue here but if we assume everything is normal then other than sat Sunday and Monday there are more count of rides so maybe this will be related to office going people

**Month**

Since Jan , Feb, Nov and dec there is snowfall in US its depicting that there are more count of rides other than these month

Q2) Why is it important to use drop_first=True during dummy variable creation?

drop_first=True is important to use, as it helps in reducing the redundant column created during dummy variable creation. When we remove the first column it takes the first column as reference for other categories. Hence it reduces the correlations created among dummy variables and ultimately helps in better model.

Q3)Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Temp and atemp has the highest correlation with the target variable

Q4)How did you validate the assumptions of Linear Regression after building the model on the training set?

1. Linearity – Pair Plot , It satisfied the assumption that linear relationship should exist
2. Mean of Residuals – using calculation for residuals(y pred – y actual) and then further mean and It passed since it was close to 0
3. Check for Homoscedasticity – Plotted residuals and visually checked the data points and it passed
4. Check for Normality of error terms/residuals – Plot distribution curve for residuals and it passed
5. No autocorrelation of residuals – checked using ***Ljungbox test*** and it failed the test , there was some correlation of residuals
6. No perfect multicollinearity- plotted corplot to check if multicollinearity exist

Q5)Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Top Features are – windspeed, Month and season

## 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 like Simple linear , Multiple Linear,

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:**

$Y = \theta_1 + \theta_2 x$

While training the model we are given :
**x:** input training data (univariate – one input variable(parameter))
**y:** labels to data (supervised learning)

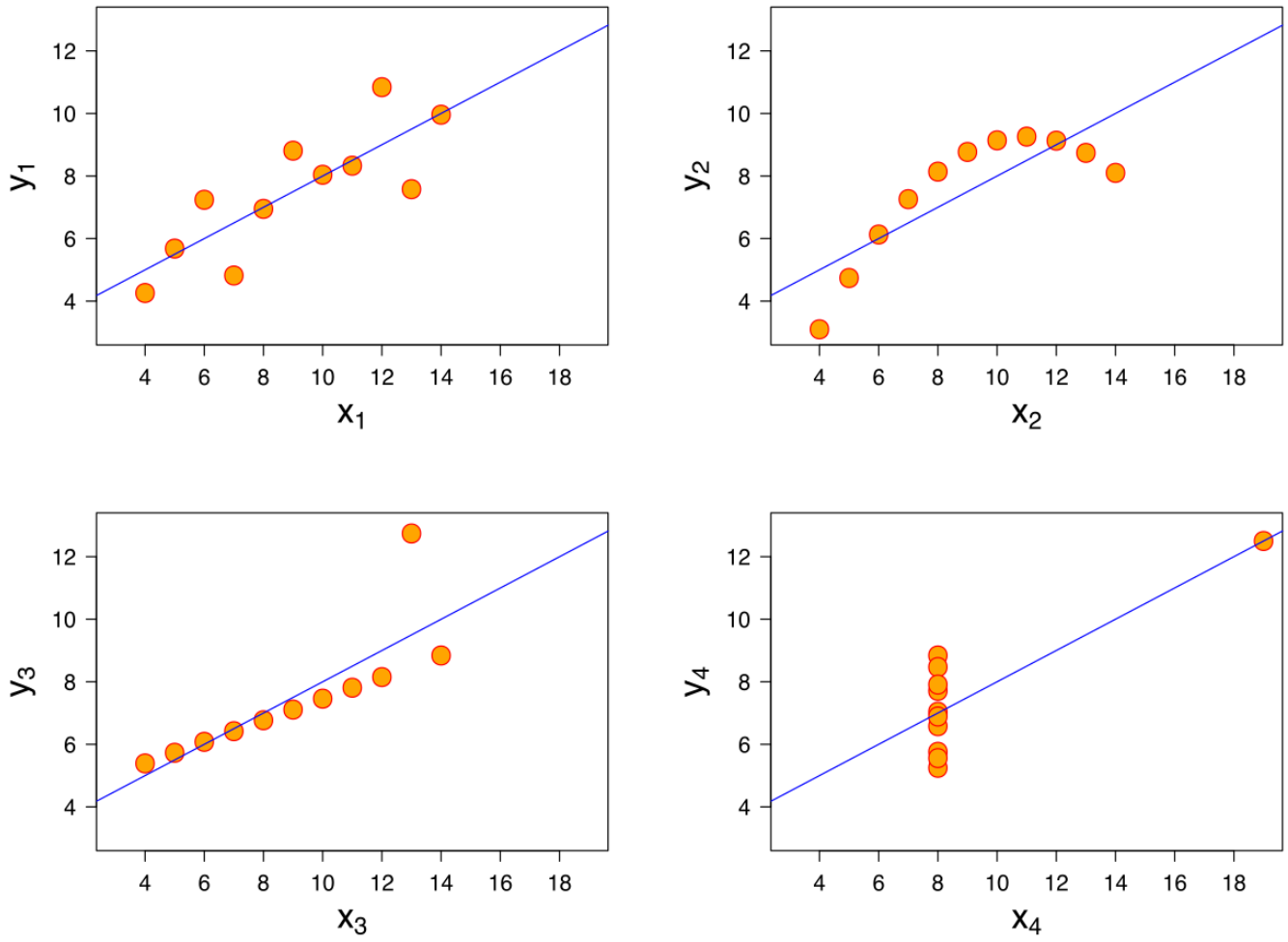
When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best $\theta_1$ and $\theta_2$ values.
**$\theta_1$:** intercept
**$\theta_2$:** coefficient of x

Once we find the best $\theta_1$ and $\theta_2$ values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

Cost Function (J):
By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the $\theta_1$ and $\theta_2$ values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y). There are various technique, includes most commonly used that is - gradient Descent, used to achieve the same. To use effectively this algorithm there are certain assumptions which are :

- Linear relationship
- Multivariate normality
- No or little multicollinearity
- No auto-correlation
- 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. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough. Below are 4 datasets graphs:




For all four datasets:

| Property | Value | Accuracy |
|---|---|---|
| Mean of x | 9 | exact |
| Sample variance of x : s2<br>x | 11 | exact |

| | | |
|---|---|---|
| Mean of y | 7.50 | to 2 decimal places |
| Sample variance of y : s2<br>y | 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<br><br>regression : {\displaystyle R^{2}} | 0.67 | to 2 decimal places |

- 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.

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.

The datasets are as follows. The x values are the same for the first three datasets

3. What is Pearson's R?

   The Pearson correlation coefficient ,also referred to as Pearson's r, is a measure of the strength and direction of linear relationships between pairs of continuous variables. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1

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

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.

**Why**?

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. Also for better interpretability also , it is required to scale features.

For example in same data set we have 2 numerical variables age and income , one can have data points around max 100 and other can go upto lacs so if scaling is not done here then model will weight more towards income feature.

Difference :-

**Min-Max Normalization:** This technique re-scales a feature or observation value with distribution value between 0 and 1. So we can see that data points will be more squeezed between 0 and 1 and there wont be any change in distribution pattern. It is a good technique to use when you do not know the distribution of your data or when you know the distribution is not Gaussian (a bell curve).

**Standardization:** 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. So here distribution will result into normal distribution. standardization is useful when your data has varying scales and the algorithm you are using does make assumptions about your data having a Gaussian distribution, such as linear regression, logistic regression, and linear discriminant analysis.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   VIF(Variance Inflation Factor) is calculated as :

$$\text{VIF} = \frac{1}{1 - R_i^2}$$

   Whenever R = 100 percent VIF would be indefinite or infinite.
   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).so whenever we find these cases , we will observe VIF as infinite

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

Use and importance in Linear Regression :

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Also it can be used to validate the assumption of normality . The normal Q Q plot is one way to assess normality,