# 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:-** There were 6 categorical variables in the dataset.

We used Box plot (refer the fig above) to study their effect on the dependent variable (‘count’)

**SEASON:**

Spring records lowest counts of bike booking and fall records the highest with a median of over 5000, so Season will be a good predictor for independent variable

**MONTH:**

we get a trand in counts as may, jun, jul, aug, sep and oct has highest medians so it can be a good predictor for independent variable

**WEATHERSIT:**

"clear weathersit" has the highest count of booking with a median close to 5000 and "light rain & snow" records rhe lowest booking count so it will be a good predictor for independent variable

**HOLIDAY:**

Bike demand is less in holidays in comparison to not being holiday.

**WEEKDAY:**

weekday variable shows very similar trends having medians closed to 5000 so variable have some or no influence on the predictor

**WORKINGDAY:**

There is no significant change in bike demand with workign day and non working day.

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

**Ans:-**

drop_first=True is important to use as it helps to reduce the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables

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

**Ans:-**

the numerical variable ‘atemp’ and "temp" has the highest correlation of 0.63 with the target variable ‘count’.

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

**Ans:-**

P-values of all the variables are very low (approximately equal to 0) and VIF values are also less than 5, which is acceptable.

F-Statistics is used for testing the overall significance of the Model: Higher the F-Statistics, more significant the Model is.

F-statistic: 242.4 Prob (F-statistic): 7.54e-167 The F-Statistics value (which is greater than 1) and the p-value of '~0.0000' states that the overall model is significant

*R-squared = 79.5 which means that 79.5 % of the variance for the target variable ie., 'count' is explained by the predictor variables , and hence we say that it is a good model.*

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

**Ans:-**

The three most significant variables affecting the demand for shared bikes are

**1.Temperature**

**2.year**

**3.weather**

# General Subjective Questions

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

Linear Regression is a Machine Learning algorithm which is a part of Supervised Machine Learning. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Based on relationship between dependent and independent variable and number of independent variable different regression models are selected. Linear regression performs the task to predict a dependent variable based on a given independent variable.

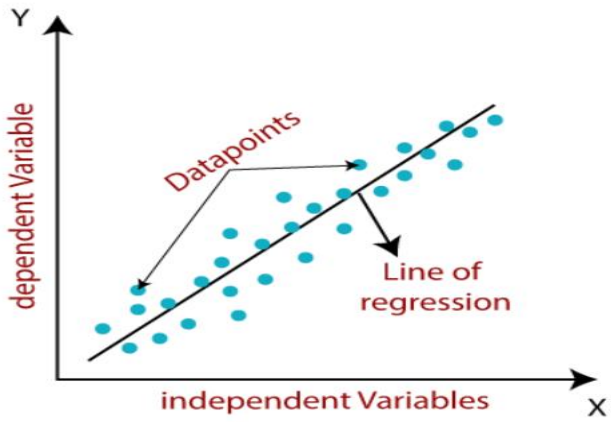
Hypothesis function for Linear Regression is Y= θ1 + θ2.x

θ1: intercept

θ2: coefficient of x

x: independent variable

y: dependent variable




Cost Function: We can measure the accuracy of our hypothesis function by using a cost function. This takes an average difference of all the results of the hypothesis with inputs from x's and the actual output y's.

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}_i - y_i)^2 = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i)^2$$

Cost function of Linear Regression is the Root Mean Squared Error between predicted and true y value. Gradient Descent: We can measure the accuracy of our hypothesis function by using a cost function. This takes an average difference (actually a fancier version of an average) of all the results of the hypothesis with inputs from x's and the actual output y's…. To reduce Cost Function by updating θ1and θ2 values and to achieve Best Fit Line the model uses Gradient Descent. The Gradient Descent algorithm is:

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$$

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

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.
This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the **data with linear relationships** and is incapable of handling any other kind of datasets. These four plots can be defined as follows:

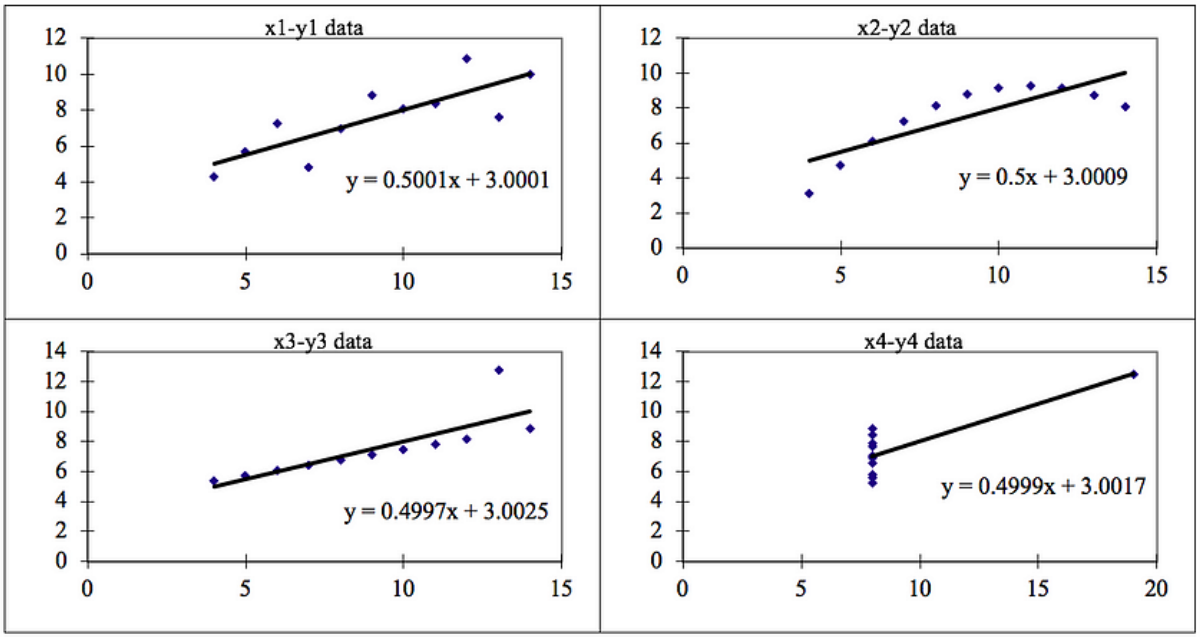
| Anscombe's Data | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Observation | x1 | y1 | | x2 | y2 | | x3 | y3 | | x4 | y4 |
| 1 | 10 | 8.04 | | 10 | 9.14 | | 10 | 7.46 | | 8 | 6.58 |
| 2 | 8 | 6.95 | | 8 | 8.14 | | 8 | 6.77 | | 8 | 5.76 |
| 3 | 13 | 7.58 | | 13 | 8.74 | | 13 | 12.74 | | 8 | 7.71 |
| 4 | 9 | 8.81 | | 9 | 8.77 | | 9 | 7.11 | | 8 | 8.84 |
| 5 | 11 | 8.33 | | 11 | 9.26 | | 11 | 7.81 | | 8 | 8.47 |
| 6 | 14 | 9.96 | | 14 | 8.1 | | 14 | 8.84 | | 8 | 7.04 |
| 7 | 6 | 7.24 | | 6 | 6.13 | | 6 | 6.08 | | 8 | 5.25 |
| 8 | 4 | 4.26 | | 4 | 3.1 | | 4 | 5.39 | | 19 | 12.5 |
| 9 | 12 | 10.84 | | 12 | 9.13 | | 12 | 8.15 | | 8 | 5.56 |
| 10 | 7 | 4.82 | | 7 | 7.26 | | 7 | 6.42 | | 8 | 7.91 |
| 11 | 5 | 5.68 | | 5 | 4.74 | | 5 | 5.73 | | 8 | 6.89 |

Image by Author

The statistical information for all these four datasets are approximately similar and can be computed as follows:

| Anscombe's Data | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Observation | x1 | y1 | | x2 | y2 | | x3 | y3 | | x4 | y4 |
| 1 | 10 | 8.04 | | 10 | 9.14 | | 10 | 7.46 | | 8 | 6.58 |
| 2 | 8 | 6.95 | | 8 | 8.14 | | 8 | 6.77 | | 8 | 5.76 |
| 3 | 13 | 7.58 | | 13 | 8.74 | | 13 | 12.74 | | 8 | 7.71 |
| 4 | 9 | 8.81 | | 9 | 8.77 | | 9 | 7.11 | | 8 | 8.84 |
| 5 | 11 | 8.33 | | 11 | 9.26 | | 11 | 7.81 | | 8 | 8.47 |
| 6 | 14 | 9.96 | | 14 | 8.1 | | 14 | 8.84 | | 8 | 7.04 |
| 7 | 6 | 7.24 | | 6 | 6.13 | | 6 | 6.08 | | 8 | 5.25 |
| 8 | 4 | 4.26 | | 4 | 3.1 | | 4 | 5.39 | | 19 | 12.5 |
| 9 | 12 | 10.84 | | 12 | 9.13 | | 12 | 8.15 | | 8 | 5.56 |
| 10 | 7 | 4.82 | | 7 | 7.26 | | 7 | 6.42 | | 8 | 7.91 |
| 11 | 5 | 5.68 | | 5 | 4.74 | | 5 | 5.73 | | 8 | 6.89 |
| | | | | Summary Statistics | | | | | | | |
| N | 11 | 11 | | 11 | 11 | | 11 | 11 | | 11 | 11 |
| mean | 9.00 | 7.50 | | 9.00 | 7.500909 | | 9.00 | 7.50 | | 9.00 | 7.50 |
| SD | 3.16 | 1.94 | | 3.16 | 1.94 | | 3.16 | 1.94 | | 3.16 | 1.94 |
| r | 0.82 | | | 0.82 | | | 0.82 | | | 0.82 | |

When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:




**Dataset 1:** this **fits** the linear regression model pretty well.
**Dataset 2:** this **could not fit** linear regression model on the data quite well as the data is non-linear.
**Dataset 3:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model
**Dataset 4:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

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

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. The Pearson's correlation coefficient varies between -1 and +1 where: r = 1 means the data is perfectly linear with a positive slope ( both variables tend to change in the same direction) r = -1 means the data is perfectly linear with a negative slope ( both variables tend to change in different directions) r = 0 means there is no linear association r > 0 < 5 means there is a weak association r > 5 < 8 means there is a moderate association r > 8 means there is a strong association As the correlation coefficient increases in

magnitude, the points become more tightly concentrated about a straight line through the data The formula of Pearson's correlation coefficient is

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$r$ = correlation coefficient

$x_i$ = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

$y_i$ = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

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

Scaling performed during the data pre-processing to handle highly varying magnitudes or values or units. It is also known as Feature Scaling which is a technique to standardize the independent features present in the data in a fixed range. Without scaling machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values. In Industry data-sets contains features highly varying in magnitude, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence we will get incorrect model. To solve this issue we have to do scaling to bring all variables to the same level of magnitude

**Normalized scaling:** It brings all the data in the range of 0 and 1 In python we use sklearn.preprocessing.MinMaxScaler to implement normalization

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization scaling: Standardization replaces the value by their Z scores. It brings all the data into a standard normal distribution which has mean (μ)=zero and standard deviation (σ)=one.

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

Sklearn.preprocessing.scale helps to implement standardization in python One advantage of normalization over standardization is that it loses some information in the data, especially about outliers. 5.You might have observed that sometime

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

**Ans:-** If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

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

**Ans:-** Q-Q plots or 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 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 two data sets come from a common distribution, the point will fall on the reference line.

If the two distribution being compared and they are similar the points in Q-Q plot will approximately lie on the line where y=x. If the distribution are linearly related to each other, the points on the Q-Q plot will approximately lie on a line but not necessarily on the line where y=x. these plots can also be used as a graphical means of estimating parameters in a location scale family of distribution. Q-Q plot is used to compare the shape of distribution, providing a graphical view that how location, scale, and skewness are similar or different in the two distributions