# Assignment-based Subjective Questions

**Question 1**. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

**Total Marks**: 3 marks (Do not edit)

**Answer:** From the categorical variables analysis we could infer following information:

- Fall and Summer Seasons (season 3 and 2) sees the highest and second highest number of bike rentals, the median bike rental starts dropping in winter and continues the downward trend during spring.
- Year 1 (2019) has seen an upward tick in bike rentals with year 1 first quartile being closer to Year 0 (2018) 3rd Quratile. We can infer that the bike sharing business was booming right before the COVID lockdown.
- We also see that bike rentals tend to go up between month 1 and 9 followed by a sharp decline during month 9-12. This can also be correlated with the Season categorical variable with warmer seasons seeing higher bike rentals while colder season have lower bike rentals. September and October sees the higher bike rentals.
- Although the IQR for weekdays wary slightly, the IQR for all weekday is quite similar. Thursday and Friday sees slightly higher rental counts.
- Non-working day have slightly higher bike rentals.
- Better weather situation are also a driving factor with clear, partly cloudy and misty days seeing higher bike rentals than rainy or snowy days.

---

**Question 2.** Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

**Total Marks:** 2 marks (Do not edit)

**Answer:** <Your answer for Question 2 goes below this line> (Do not edit)

By default, pandas get_dummies function will create n dummy variables for any variable with n categories, however we only need n-1 variables to define n categories. Dropping one of the dummy variable also help in reducing the number of overall predictor variables.

---

**Question 3.** Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

**Total Marks:** 1 mark (Do not edit)

**Answer:** <Your answer for Question 3 goes below this line> (Do not edit)

Temp variable has the highest correlation with cnt (target variable)

---

**Question 4.** How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** <Your answer for Question 4 goes below this line> (Do not edit)

Following steps were taken to validate the assumptions of Linear Regression:

1. **Residual Analysis**

We calculated the residuals or error terms for the predicted values. These values were then plotted on a distplot which confirm following assumptions.

- The mean of errors is centered around 0 and the errors are normally distributed..

2. **Linear Relationship between predictors and target variables**

This was taken care of by making sure that all the predictor variables are significant ie have a p-value less than 0.05. The R-Squared and Adjusted R-Squared values are .798 and .794 this means that variance in data is being explained by all the selected variables.

### 3. Predictor Variables are Independent of each other

All selected predictor variables have a VIF of less than 5 meaning all variables are independent of each other and there is no multi-collinearity issue.

---

**Question 5.** Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
**Total Marks:** 2 marks (Do not edit)
**Answer:** <Your answer for Question 5 goes below this line> (Do not edit)
Top 3 features contributing to bike rentals are

- **Temp** (coef : **4405.28**)
- **Yr** (coef:**2054.3966**)
- **Good**(weathersit 1) (coef:**827.4753**)

---

# General Subjective Questions

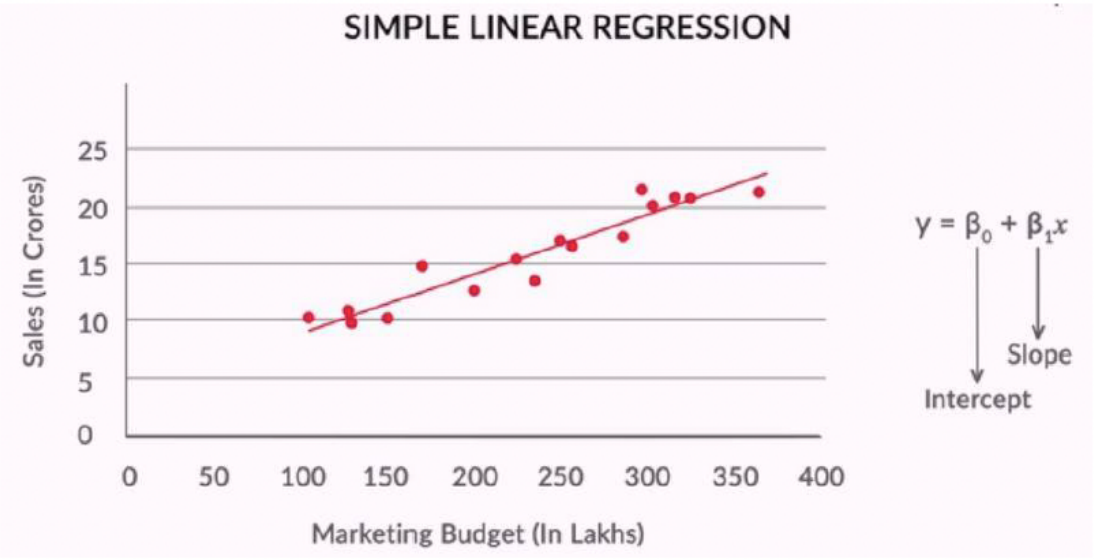
**Question 6.** Explain the linear regression algorithm in detail. (Do not edit)
**Total Marks:** 4 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

Linear Regression is a statistical method that is used to model the relationship between on or more independent variables called predictor variables, and one target variable called predicted/calculated variable. This model tries to find the best fitting line for the given data. This line can then be used to find the relationship between variables, predict deviations and target values for given predictor.

**Linear Regression Types:**
**Simple-Linear Regression-**
Simple Linear Regression tries to predict the target variable using one predictor variable. The movel is represented with equation **y = mx+b**.

**Multiple Linear Regression**
Multiple Linear regression involves multiple predictor variables and tries to predict the value of a single target variable. The model is represented by **y = b0 + b1x1 + b2x2 + …… bnxn.**

**Linear Regression Concepts:**
Linear Regression can be represented with following figure.




Based on the figure following are the key parameters involved in Linear Regression.

1. **Intercept** (b0 or m) – variable coefficient or slope of the line represent the change in target variable with unit changes in the predictor variable if all other variables are constant.
2. **Y (Target Variable) –** This is the variable that we want to predict.
3. **X (independent variable) –** This is the variable on which the target variable depends, this variable influences the target variable value.
4. **b1(intercept) –** this is the value of the target variable if all the predictor variables are zero.
5. **Residual or error terms–** Residual or error is the difference between the actual target value and the predicted target value.

**Linear Regression Assumptions –**

Linear Regression is based on following assumptions –

1. It is assumed that there is a linear relationship between the predictor variables and target variables.
2. Independent variables are measured without errors.
3. It is also assumed that the predictor variables are independent of each other and do not have high correlation with each other.
4. It is assumed that predictions are independent of each other.
5. Residual Assumptions
    a. It is assumed that the mean of residuals or errors terms is zero.
    b. It is also assumed that the error terms are normally distributed and centered around zero.
    c. The variance of errors is assumed to be constant across all levels of predictor variables.
    d. Residuals are assumed to be independent of each other's and do not follow or form a pattern.

**Question 7.** Explain the Anscombe's quartet in detail. (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

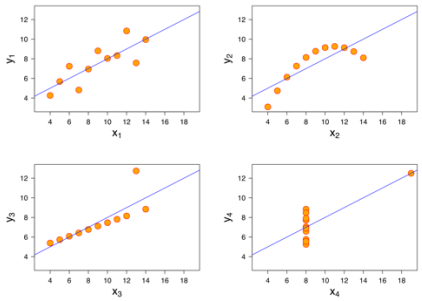
Francis Anscombe in 1973 created a set of 4 dataset which have nearly identical summary statistics however have completely different visual patterns. Anscombe's quartet focuses on the importance of visualizing the dataset and not solely rely on the summary statistics.

The 4 datasets consists of 11 datapoints and have following summary statistics.

| Property | Value | Accuracy |
|---|---|---|
| Mean of $x$ | 9 | exact |
| Sample variance of $x$: $s_x^2$ | 11 | exact |
| Mean of $y$ | 7.50 | to 2 decimal places |
| Sample variance of $y$: $s_y^2$ | 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 regression: $R^2$ | 0.67 | to 2 decimal places |

---

The summary statistic might lead one to believe that all four datasets are quite similar in nature, however when plotted all four turn out to be considerably different.




As per the plots:

1. The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two correlated variables, where y could be modelled as gaussian with mean linearly dependent on x.
2. For the second graph (top right), 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.
3. In the third graph (bottom left), the modelled relationship 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.
4. 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.

---

**Takeaway for Anscombe's Quartet**

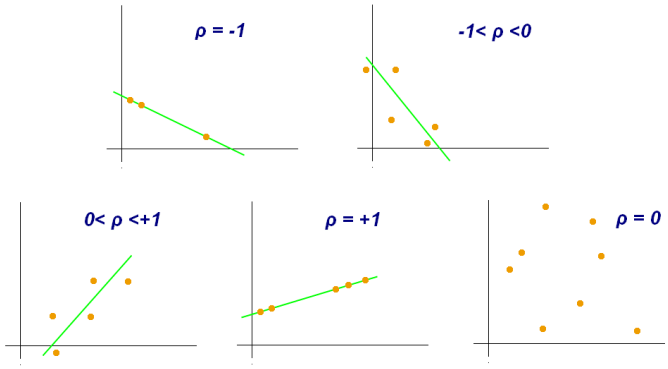
Anscombe's quartet serves as a powerful reminder of the importance of data visualization in understanding the underlying patterns and relationships within a dataset. In some cases, the summary statistics might be misleading, and visual representation might reveal completely different relationships then what's presented by summary statistics.

---

**Question 8.** What is Pearson's R? (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

In Statistics, Pearson's R or Pearson Correlation Coefficient (r) is one of the most common ways of measuring a linear correlation. It varies between -1 and 1. The value measures the strength and direction of the relationship (positive or negative) between two variables.

It is the ratio between the covariance of two variables and 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. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationships or correlations.




**Key points to remember:**

1. The value ranges from -1 to 1.
2. Positive values mean positive linear correlation meaning if one variable increases the other variable also increases.
3. Negative values mean negative linear correlation meaning if one variable increases the other variable will decrease.
4. The value itself denotes the strength of relationship with 1 and -1 depicting perfect positive and perfect negative relationship.
5. 0 denotes no linear relationship between the variables.

---

**Question 9.** What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

When performing data analysis, scaling refers to the process of transforming the variables in such a way that all the involved variables are on the same scale.

When working with real world data, the data range might vary widely, for example- let's say we are trying to calculate the pattern in fuel economy of different cars based on distance travelled, engine temperature, engine RPMs and average speed. In this dataset while the engine temperature will have smaller values, the engine RPMs will be in thousands and can dominate other variables.

Many of the machine learning algorithms work by calculating the distances between two points by the Euclidean distance, if all datapoints are not on the same scale then the distance will be governed by the larger variables and other variables will become insignificant.

Another reason is that scaling can improve the convergence speed and stability of the Gradient Descent based algorithms such as Linear and Logistic Regression.

**Normalized Scaling –** Normalized scaling is also known as MinMax Scaling, it transforms the variable values between 0 and 1. It is one of the simplest methods to implement and preserves the original data range.

---

**Formula –**

```
x` = (x - min(x)) / (max(x) - min(x))
```

**Standardized Scaling -** Standardized Scaling transforms the features in such a way that they have zero mean and unit variance. This is done by calculating the distribution mean and standard variation for each feature and then by calculating the new datapoint.

**Formula -**

x` = (x - mean(x)) / std(x)

---

**Question 10.** You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

VIF (Variance Inflation Factor) is a predictor of degree of multicollinearity for a give predictor variable meaning it predicts how effectively one predictor can be perfectly predicted from other predictor variables.

VIF is calculated as
VIF(i) = 1/(1-Ri2)

In cases when Ri2 is 1, meaning the variable can be perfectly predicted by the other variables, the denominator of VIF formula becomes 0 resulting in infinite value for VIF.

An infinite VIF indicates perfect multicollinearity.

---

**Question 11.** What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
(Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

Q-Q plot or (Quantile-Quantile) plot is a graphical representation which is used to compare two qunatiles against each other. A quantile is a fraction where certain values lie below and other lie above that value. For example, median represents a quantile where half of the values are below that data point while the other half are above that data point.

Q-Q plot is used to figure out if both dataset come from the same distribution. If both sets of qunatiles come from the same distribution then we should see the points forming a line that;s roughly straight.

Q-Q plot can be used in normal distribution validation, **Linear Regression analysis** is based on the assumption that the data is normally distributed, Q-Q plot can be used as a visual tool to check this assumption, if the data points deviate significantly from a straight line that the assumption about the normal distribution might be violated.