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

   The insights from the box plots with categorical variables and count variable are as follows:
   - In fall season, the rentals are higher
   - The rentals have increased from 2018 to 2019
   - Months: Jul to Sep, the rentals are higher
   - When weather is clear, the rentals are higher

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

   For Categorical variable with k levels, we would need k-1 columns to represent the data in dummy variables.

   When we use **get_dummies** function for a variable with k levels, it would return k columns. So, **drop_first = True** will drop the first column thereby reducing the columns to k-1 which is desired.

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

   '**temp'** and '**atemp'** variables have the highest correction with count variable which is around 0.63

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   - Plotted scatter plot between dependent and independent variables and observed linearity relationship
   - Plotted error terms and observed a normal distribution of error terms with mean zero.
   - Plotted scatter plot to see for any patterns in the residuals and found no patterns
   - Checked for multi collinearity using VIF and all the variables in the final model had VIF < 5

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   - Temp – a unit increase in temp increases the rentals by 0.3937
   - Year – year 2019 increases the rentals by 0.2356
   - September – a unit increase in month September increases the rentals by 0.0531

1. Explain the linear regression algorithm in detail.

Linear regression is one of the predictive modelling regression techniques in which the dependent(target) variable has a linear relationship with the independent(predictor) variables.

There are two types of linear regressions – Simple Linear Regression and Multiple Linear Regression.

- Simple Linear Regression – when the number of independent variables is 1.
- Multiple Linear Regression – when the number of independent variables is more than 1

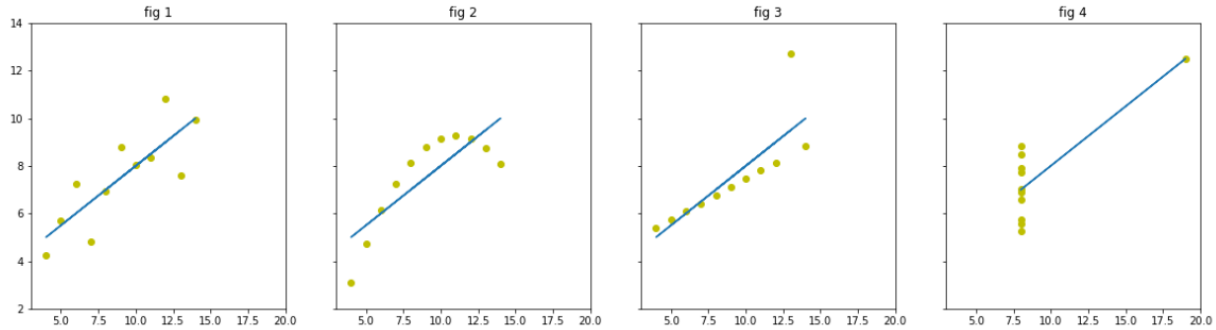
- Standard equation of the regression line is given by the equation:
  - $y = \beta 0 + \beta 1\ X1 + \beta 2\ X2 + \ldots\ldots + \beta n\ Xn$
- The best-fit line is found by minimizing the Residual Sum of Squares(RSS) using Ordinary Least Squares method.
- The strength of the linear regression model can be assessed by metrics like
  - $R^2 = 1 - (RSS/TSS)$ or Residual Standard Error (RSE).
- Examples where linear regression can be used - predicting the housing prices, understanding relationship between advertising spending and revenue generated etc.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises of 4 datasets that have identical statistical properties but when plotted, the graphs are different. Each dataset consists of 11 (x,y) data points.

- The mean of x1, x2, x3, x4 = 9.0 and mean of y1, y2, y3, y4 = 7.5
- Slope of the line for all datasets = 0.5
- Intercept of the line for all datasets = 3.0

But the graphs of the 4 datasets when plotted will look like below.



So, it is always recommended to visualize any dataset.

3. What is Pearson's R?

Pearson correlation coefficient R is a measure of the strength of the linear relationship between two variables. Pearson's R can range from – 1 to 1. R = -1 indicates perfect negative correlation and R=1 indicates perfect positive correlation.

In general,

- If one variable increases as the other increases, then the correlation is positive
- If one variable increases as the other decreases, then the correlation is negative
- If one variable is constant as the other increases or decreases, then the correlation is zero.

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

Scaling is a pre-processing step which is applied to the independent variables to normalize the data within a particular range.

Scaling is done for:

- Ease of interpretation when the variables are at the same scale
- Faster convergence for Gradient descent methods.

Scaling just affects the coefficients and doesn't impact the parameters like p-value, R^2, T-statistic or F-statistic.

Normalized scaling: Normalization technique brings all the data between the range 0 and 1. It is also known as Min-Max scaling. Given by the equation

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

Standardized scaling: Standardization technique brings all the data into a standard normal distribution with mean 0 and standard deviation 1. Given by the equation

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

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

VIF is calculated by the formula $VIF = 1/(1-R^2)$. If R^2 value is 1, then the above formula will give the value of VIF as infinity. We get $R^2 = 1$ in case of perfect correlation.

To solve this problem, one of the variables needs to be dropped from the dataset which causes this perfect collinearity.

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

Q-Q plot is a scatterplot created when two sets of quantiles are plotted against each other. If two datasets come from a population with the same distribution, the data points would fall approximately along the 45-degree reference line which is plotted.
In general, it helps in answering the following:

- If two datasets come from populations with a common distribution
- If two datasets have a similar distribution shape
- If two datasets have similar tail behavior
- If two datasets have common location and scale

Whenever there are two data sets, it is useful to know if the assumption of the common distribution is justified. The Q-Q plot can provide more insights into the nature of the difference between data sets.