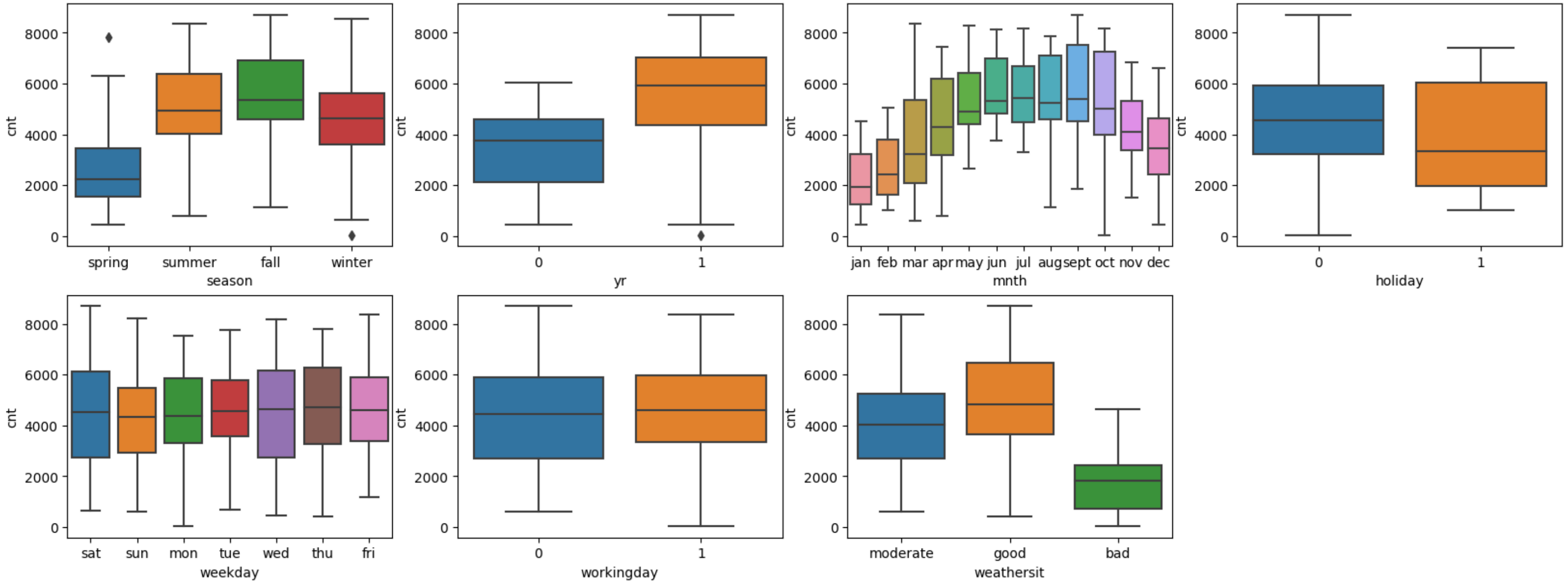
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

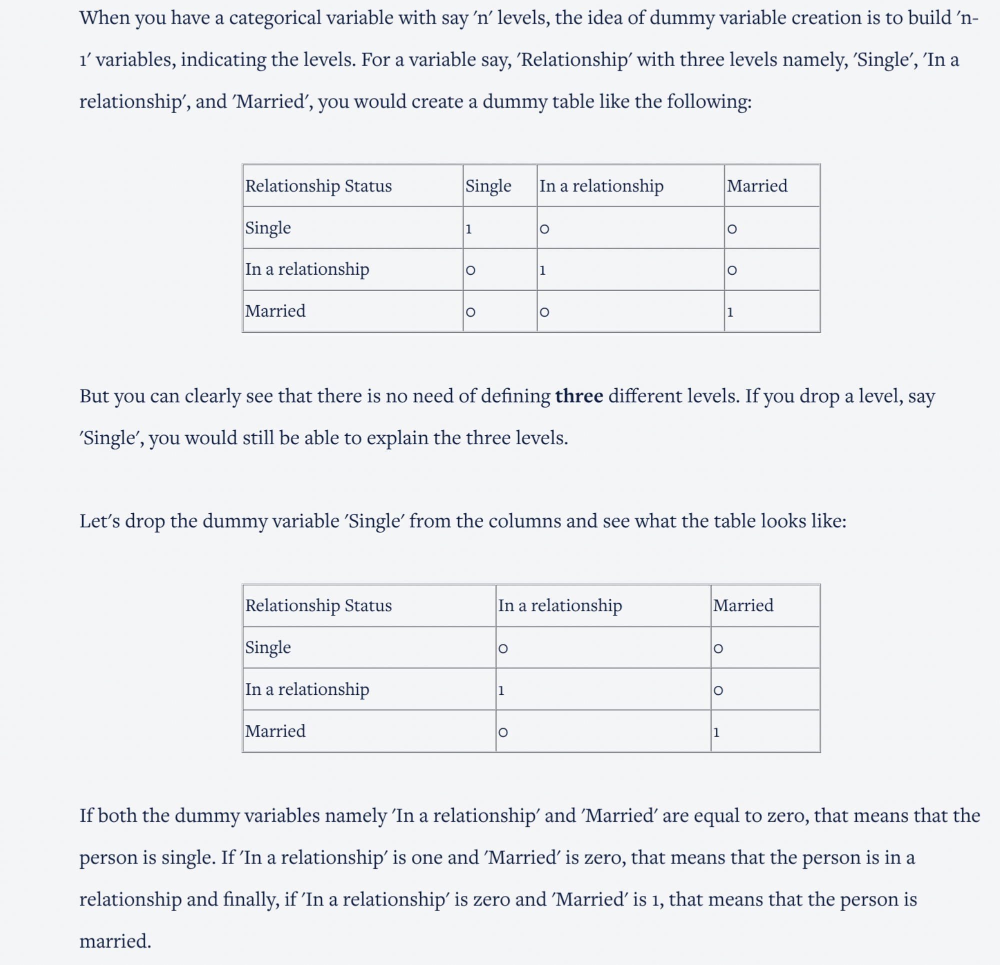
> **Ans:**
>
> 
>
>
> - **Season variable :** Fall has Highest bookings, Customer's prefer bike during Fall and summer compared winter and spring because of its climatic conditions.
> - **Yr variable :** Year on Year bookings are increasing. Customers like bike sharing concept.
> - **Month variable :** Bookings are increasing continously from jan till june and reduced a little , may be due to peak summer and again increased till sept. After Sept due to cold and rainy weather bookings reduced.
> - **Holiday variable :** On holidays customer's prefer taking rest than to commute.
> - **Weekday variable :** Cusomer's book on alternate days of week, may be because of Hybrid work model.
> - **Working Day variable :** Though median is almost same for working and non-working day. Customers using bikes for all purposes Daily commutes and weekend shoppings..

> - **Weather variable :** Customer's use bikes when they are no rains, and no heavy snowfall. It is evident with results.

---

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

> **Ans:** It is important to use drop_first=True during dummy variable creation because , it reduces the extra column to created while creating dummy variables. If we have N levels of values in categorical variable we can explain the variable with N-1 levels also. See below for Example.



When you have a categorical variable with say 'n' levels, the idea of dummy variable creation is to build 'n-1' variables, indicating the levels. For a variable say, 'Relationship' with three levels namely, 'Single', 'In a relationship', and 'Married', you would create a dummy table like the following:

Relationship Status	Single	In a relationship	Married
Single	1	0	0
In a relationship	0	1	0
Married	0	0	1

But you can clearly see that there is no need of defining **three** different levels. If you drop a level, say 'Single', you would still be able to explain the three levels.

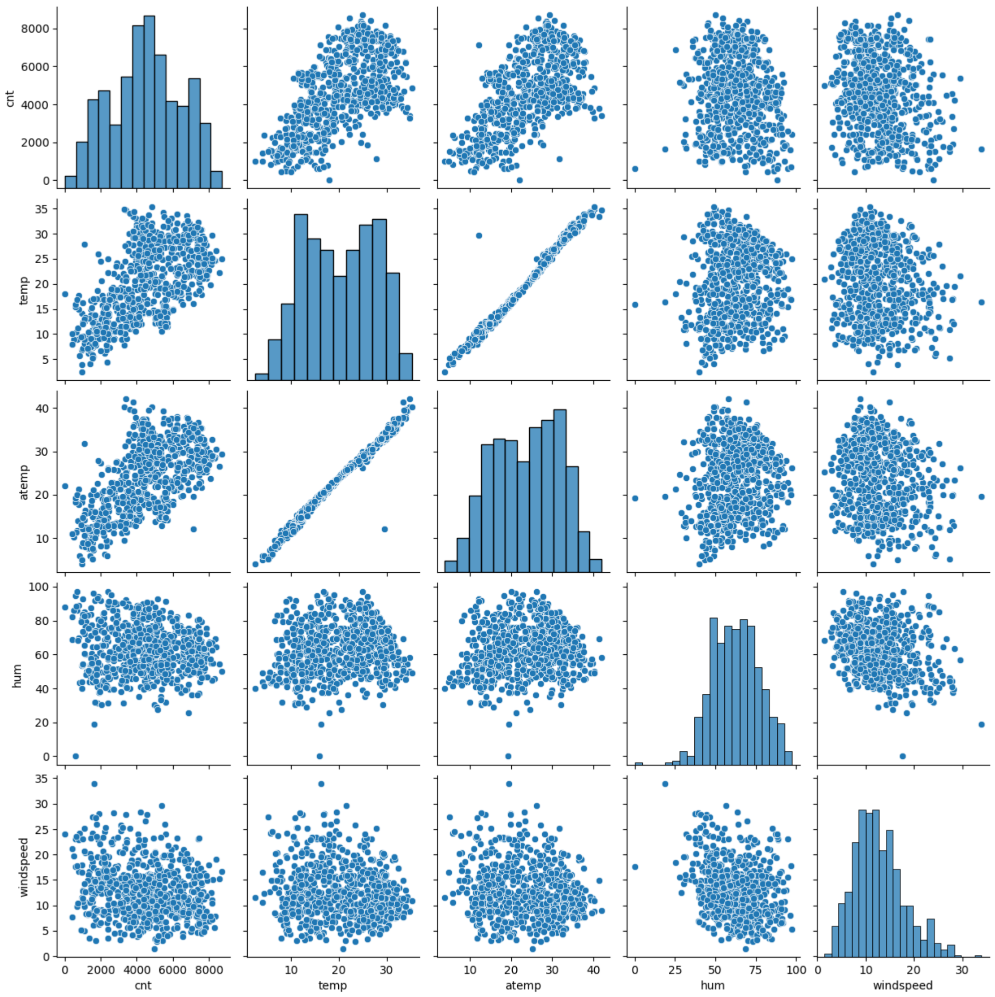
Let's drop the dummy variable 'Single' from the columns and see what the table looks like:

Relationship Status	In a relationship	Married
Single	0	0
In a relationship	1	0
Married	0	1

If both the dummy variables namely 'In a relationship' and 'Married' are equal to zero, that means that the person is single. If 'In a relationship' is one and 'Married' is zero, that means that the person is in a relationship and finally, if 'In a relationship' is zero and 'Married' is 1, that means that the person is married.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

> **Ans:** After analysing pair-plot among the numerical variables, cnt(target variable) has highest correleation with temp and atemp variables. Moreover temp and atemp has clear linear relationship, So we can remove any one of variable as it is redundant in anylsis after obtaining VIF and P-Value. See below for reference of Pair-plot.



4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

> **Ans:**
>
> > 1. **Linear Relationship:** linear regression assumes that there exists a linear relationship between the dependent variable and the predictors.
> >
> >   - *Verification:* Pair-wise scatterplots will be helpful in validating the linearity assumption as it is easy to visualize a linear relationship on a plot.
> > 2. **Homoscedasticity:** Homoscedasticity means that the residuals have constant variance no matter the level of the dependent variable.
> >
> >   - *Verification:* By looking at the residual plot and we can verify that the variance of the error terms is constant across the values of the dependent variable.
> > 3. **Absence of Multicollinearity:** Multicollinearity refers to the fact that two or more independent variables are highly correlated (or even redundant in the extreme case). While it may not be important for non-parametric methods, it is primordial for parametric models such as linear regression.
> >
> >   - *Verification:* First using correlation Heat Map of variables and second by calculating VIF (Variance Inflation Factors)
> > 4. **Normality of Errors:** Error Terms should be normally Distrubuted and have mean 0
> >
> >   - *Verification:* Plot a distplot using error terms.
> > 5. **Check R2 Value:** R2 Value should be greater 75%.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

> **Ans:** Below are Three top features contributing towards demand of shared bikes

1. **Temperature:** Temp has highest positive correlation, which mean's demand of bikes increases as Temperature Increases.

2. **Year:** Demand for bike's increases year on year.

3. **Sept Month:** Sept month has highest number bookings

Final Model Results

```
build_model(significant_cols)
```

3] ✓ 0.4s

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.789
Model:                            OLS   Adj. R-squared:                  0.785
Method:                 Least Squares   F-statistic:                     207.2
Date:                Tue, 13 Dec 2022   Prob (F-statistic):          1.85e-162
Time:                        18:55:43   Log-Likelihood:                 434.76
No. Observations:                 510   AIC:                            -849.5
Df Residuals:                     500   BIC:                            -807.2
Df Model:                           9                                         
Covariance Type:            nonrobust                                         
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                   0.2429      0.027      9.041      0.000       0.190       0.296
yr                      0.2397      0.009     25.766      0.000       0.221       0.258
holiday                -0.0851      0.029     -2.885      0.004      -0.143      -0.027
temp                    0.4612      0.034     13.424      0.000       0.394       0.529
windspeed              -0.1687      0.028     -6.000      0.000      -0.224      -0.113
season_spring          -0.1064      0.017     -6.211      0.000      -0.140      -0.073
season_winter           0.0351      0.014      2.520      0.012       0.008       0.062
mnth_jul               -0.0767      0.020     -3.906      0.000      -0.115      -0.038
mnth_sept               0.0468      0.018      2.630      0.009       0.012       0.082
weathersit_moderate    -0.0677      0.010     -6.881      0.000      -0.087      -0.048
==============================================================================
Omnibus:                      133.771   Durbin-Watson:                   1.991
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              487.648
Skew:                          -1.164   Prob(JB):                    1.28e-106
Kurtosis:                       7.187   Cond. No.                         14.0
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

<statsmodels.regression.linear_model.RegressionResultsWrapper at 0x29f0eea40>
```

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

> **Linear Regression** is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e it finds the linear relationship between the dependent and independent variable.
>
> - Linear Regression is of two types: Single and Multiple.
>
>   > 1. Single Linear Regression is where only one independent variable is present and the model has to find the linear relationship of it with the dependent variable
>   >
>   > 2. Multiple Linear Regression there are more than one independent variables for the model to find the relationship.
>
> Equation of Simple Linear Regression, where bo is the intercept, b1 is coefficient or slope, x is the independent variable and y is the dependent variable.
>
> > `y= b0 + b1x`
>
> Equation of Multiple Linear Regression, where bo is the intercept, b1,b2,b3,b4…,bn are coefficients or slopes of the independent variables x1,x2,x3,x4…,xn and y is the dependent variable.
>
> > `y=b0 + b1x1 + b2x2 + b3x3 + b4x4 ... + bnxn`
>
> **A Linear Regression model's main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized.**
>
> **Assumptions of Linear Regression**

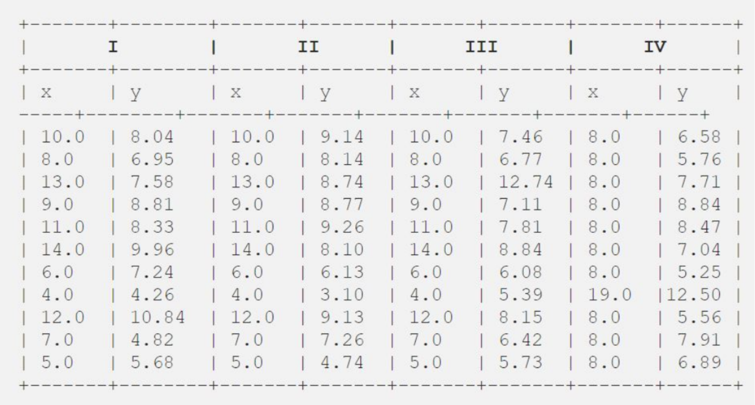
> > 1. **Linearity:** It states that the dependent variable Y should be linearly related to independent variables. This assumption can be checked by plotting a scatter plot between both variables.
> >
> > 2. **Normality:** The X and Y variables should be normally distributed. Histograms, KDE plots, Q-Q plots can be used to check the Normality assumption.
> >
> > 3. **Homoscedasticity:** The variance of the error terms should be constant i.e the spread of residuals should be constant for all values of X. This assumption can be checked by plotting a residual plot. If the assumption is violated then the points will form a funnel shape otherwise they will be constant.
> >
> > 4. **Independence/No Multicollinearity:** The variables should be independent of each other i.e no correlation should be there between the independent variables. To check the assumption, we can use a correlation matrix or VIF score. If the VIF score is greater than 5 then the variables are highly correlated.
> >
> > 5. The **error terms should be normally distributed**. Q-Q plots and Histograms can be used to check the distribution of error terms.
> >
> > 6. **No Autocorrelation:** The error terms should be independent of each other. Autocorrelation can be tested using the Durbin Watson test. The null hypothesis assumes that there is no autocorrelation. The value of the test lies between 0 to 4. If the value of the test is 2 then there is no autocorrelation.
>
> **Evaluation Metrics for Regression Analysis**
>
> To understand the performance of the Regression model performing model evaluation is necessary. Some of the Evaluation metrics used for Regression analysis are:
>
> > 1. **R squared or Coefficient of Determination:** The most commonly used metric for model evaluation in regression analysis is R squared. It can be defined as a Ratio of variation to the Total Variation. The value of R squared lies between 0 to 1, the value closer to 1 the better the model.
> >
> > 2. **Adjusted R squared: It is the improvement to R squared.** The problem/drawback with R2 is that as the features increase, the value of R2 also increases which gives the illusion of a good model. So the Adjusted R2 solves the drawback of R2. It only considers the features which are important for the model and shows the real improvement of the model. Adjusted R2 is always lower than R2.

> > 3. **Mean Squared Error (MSE):** Another Common metric for evaluation is Mean squared error which is the mean of the squared difference of actual vs predicted values.
> >
> > 4. **Root Mean Squared Error (RMSE):** It is the root of MSE i.e Root of the mean difference of Actual and Predicted values. RMSE penalizes the large errors whereas MSE doesn't.

---

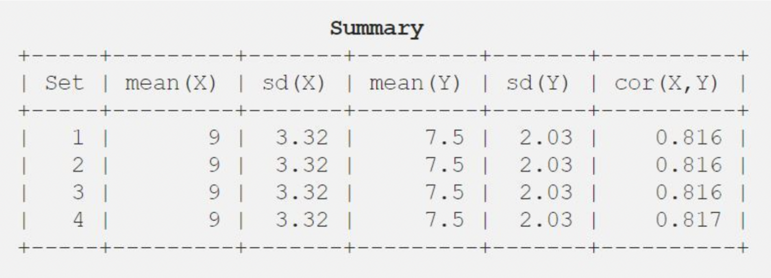
2. Explain the Anscombe's quartet in detail. (3 marks)

> Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet 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 on statistical properties.
>
> `Simple understanding:`
> Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points.
>
> Those 4 sets of 11 data-points are given below.



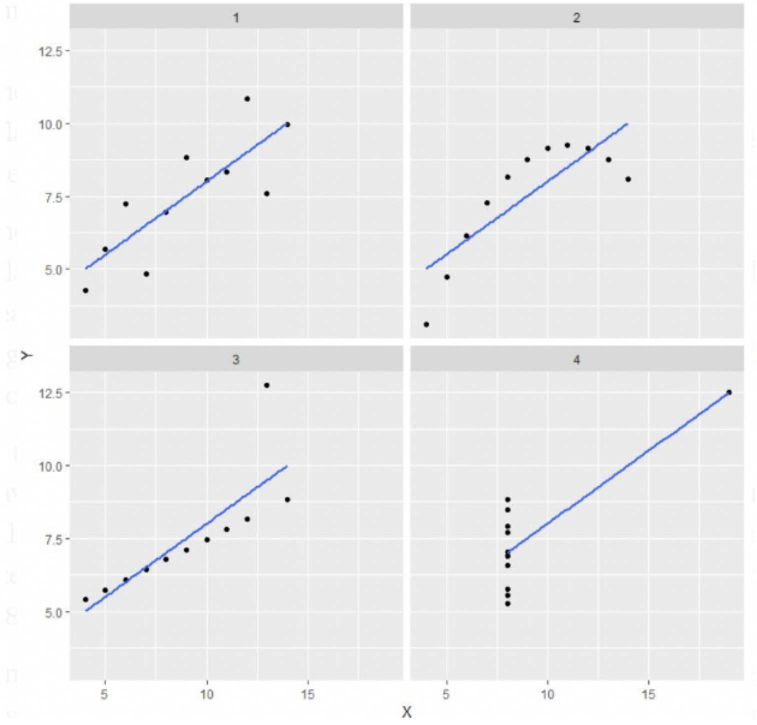
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

> After that, the council analyzed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y. result in a tabular fashion for better understanding.



```
                       Summary
+-----+---------+-------+---------+-------+----------+
| Set | mean(X) | sd(X) | mean(Y) | sd(Y) | cor(X,Y) |
+-----+---------+-------+---------+-------+----------+
|   1 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   2 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   3 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   4 |       9 |  3.32 |     7.5 |  2.03 |    0.817 |
+-----+---------+-------+---------+-------+----------+
```

Let's Visualize and interpret these values Note: It is mentioned in the definition that Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.



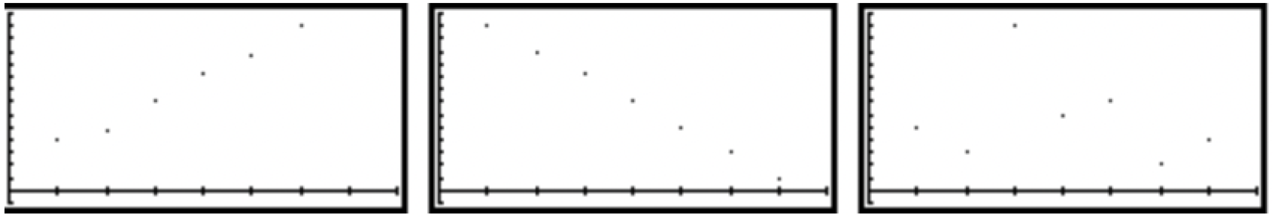
Explanation of this output:

- In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.

> - Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.
>
> **Application:** *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? (3 marks)

> 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 (i.e., both variables tend to change in the same direction)
> - r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
> - r = 0 means there is no linear association
>
> The figure below shows some data sets and their correlation coefficients. The first data set has an r=0.996, the second has an r = -0.999 and the third has an r= -0.233
>
> > 

---

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

> **Scaling** is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
>
> Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
>
> It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.
>
> > ## Difference between normalized scaling and standardized scaling
> >
> > The difference between them lies in the way they calculate and values normalization range.
> >
> > 1. Normalization/Min-Max Scaling:
> >
> > > - It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.
> >
> > 2. Standardization Scaling:
> >
> > > - Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

---

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

> If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

> 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)

---

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

> 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.
>
> 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.
>
> **Advantages:**
>
> - It can be used with sample sizes also
> - Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.
>
> **It is used to check following scenarios: If two data sets —**
>
> 1. come from populations with a common distribution
> 2. have common location and scale
> 3. have similar distributional shapes
> 4. have similar tail behavior