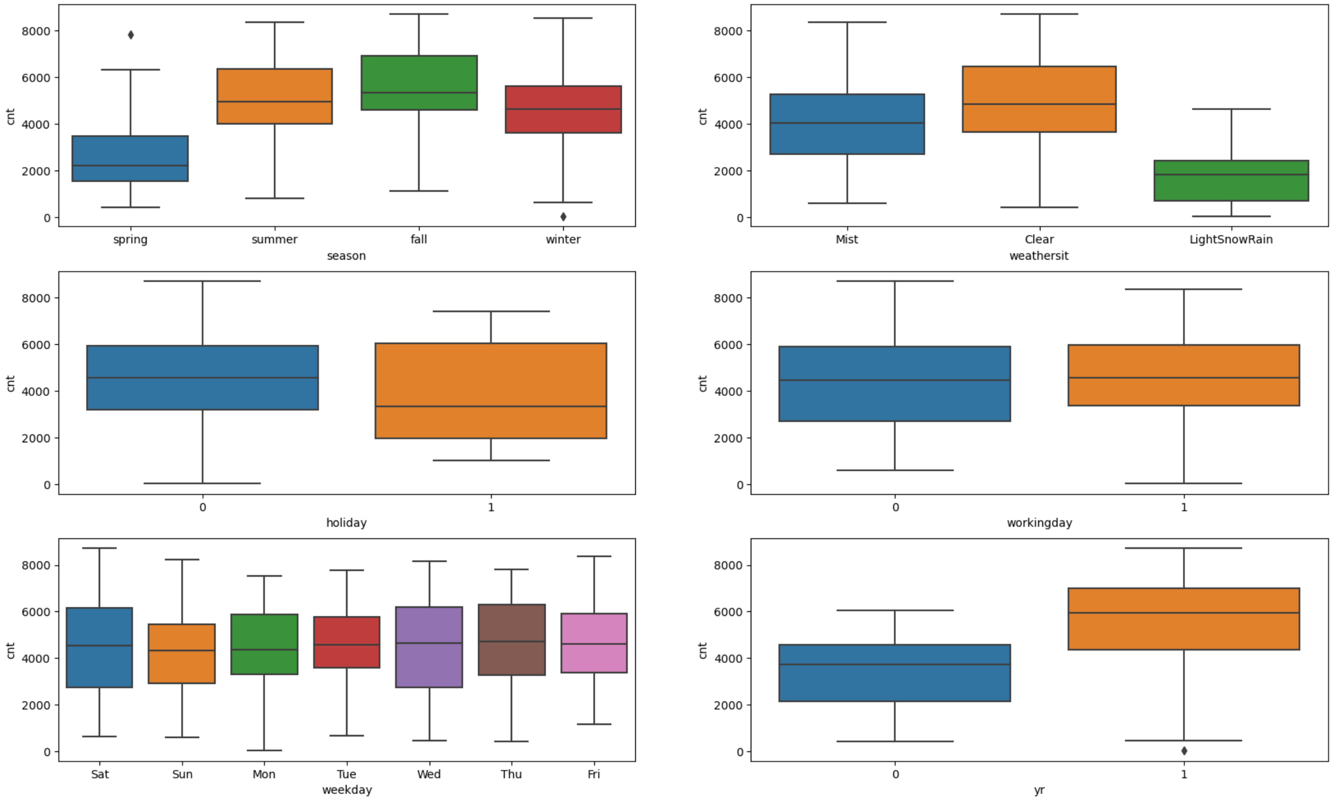
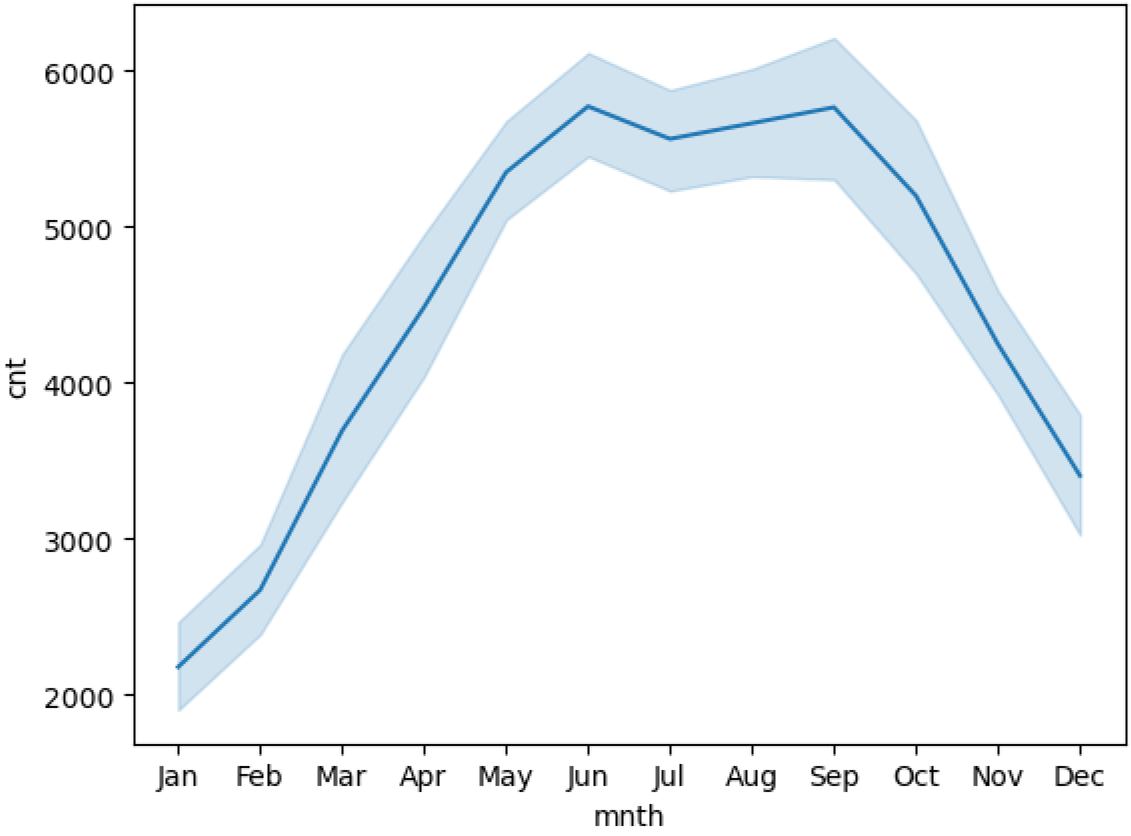
Linear Regression Subjective Questions

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?





Key Observations

- `season` - More demand is observed in 'fall' followed by 'summer' and 'winter'. Least is observed during spring
- `weathersit` - More demand is observed when the weather is 'clear', followed by 'Mist'. Least is observed during 'Light Snow / Rain'. However there seems to be no data when the weather is extreme like 'Heavy Rain' etc. Assuming the demand is zero or close to zero in such extreme weather

- `holiday` - Median demand is actually high when it is not a holiday. On holidays, it is observed that the minimum demand can go subsequently low when compared with not a holiday.
- `workingday` - Median is almost same but the minimum demand is high in case of working days.
- `weekday` - Median is almost comparable on all the days of the week. Peak demand is observed during Saturdays and Wednesdays and Thursdays. At the same time minimum demand was also observed on Saturday, Wednesday and Sunday.
- `yr` - Year 2019 seems to have got more demand. Even the minimum of 2019 is comparable to the maximum of 2018. May be because this company got barely started in 2018.
- `mnth` - This is also in sync with season and weather conditions. Demand is more between 'May' and 'Sep' months. Demand is very less in the months of Jan, Feb and Mar months

Question 2: Why is it important to use `drop_first=True` during dummy variable creation?

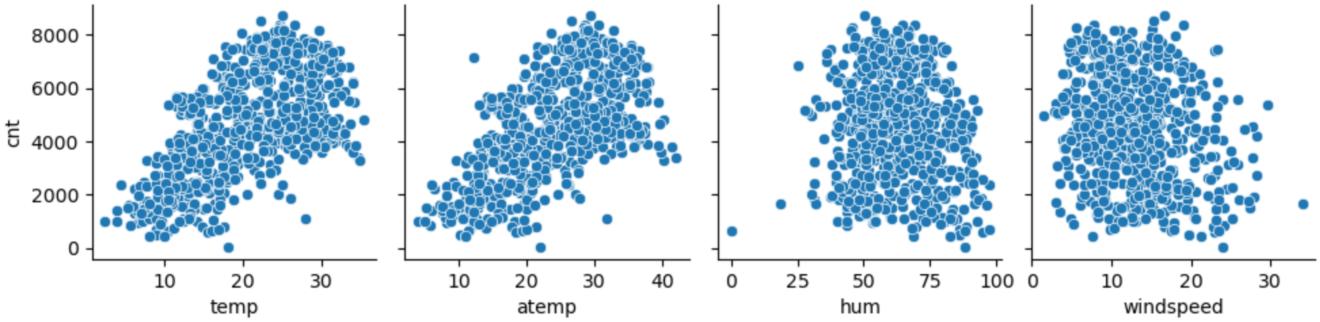
Answer:

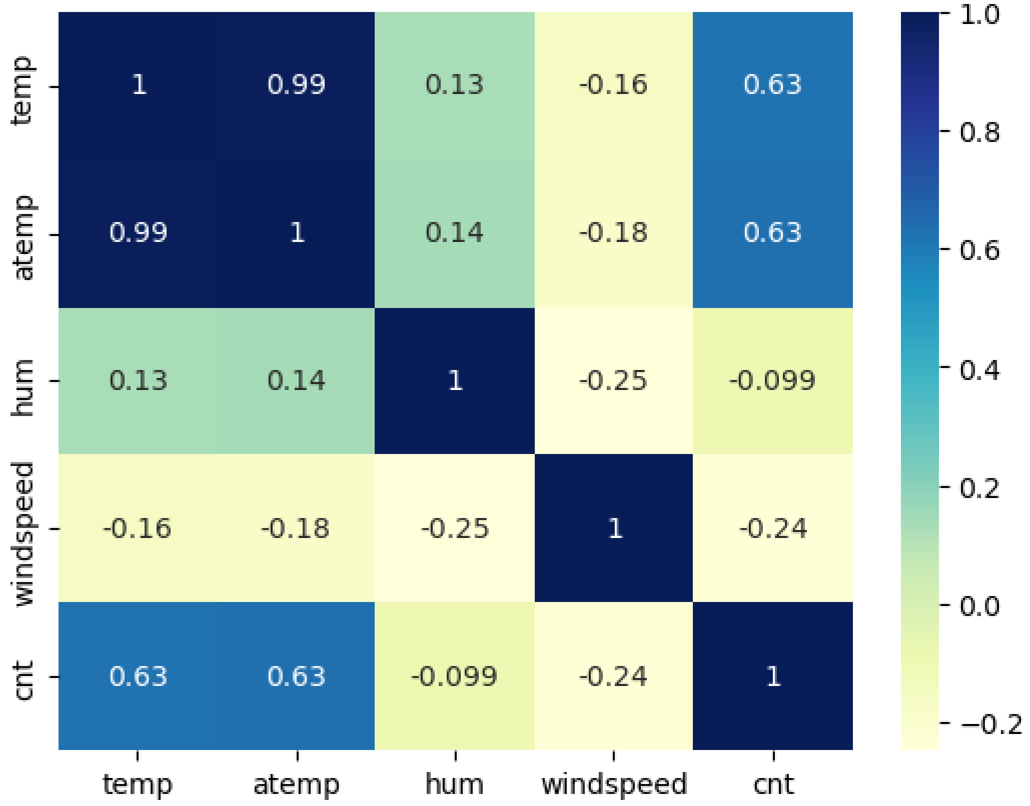
- This is basically to avoid multi collinearity as including all dummy variables can create perfect correlation and it can lead to unstable coefficient estimates.
- Typically we create only N-1 dummies as the value of the other dummy variable can be inferred from the values of N-1 dummies.

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

Answer:

- `temp` and `atemp` seem to have strong linear relationship with target variable `cnt` and it is good enough
- At the same time, `temp` and `atemp` are highly collinear

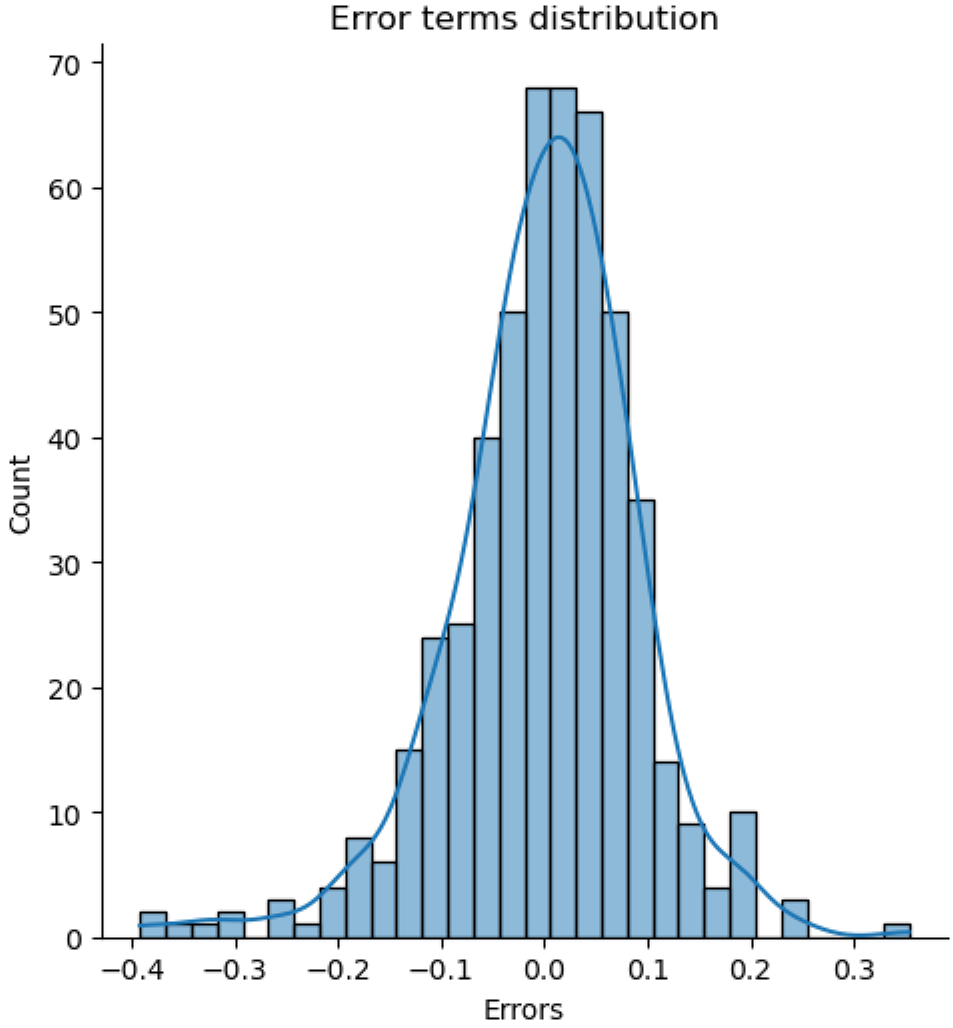


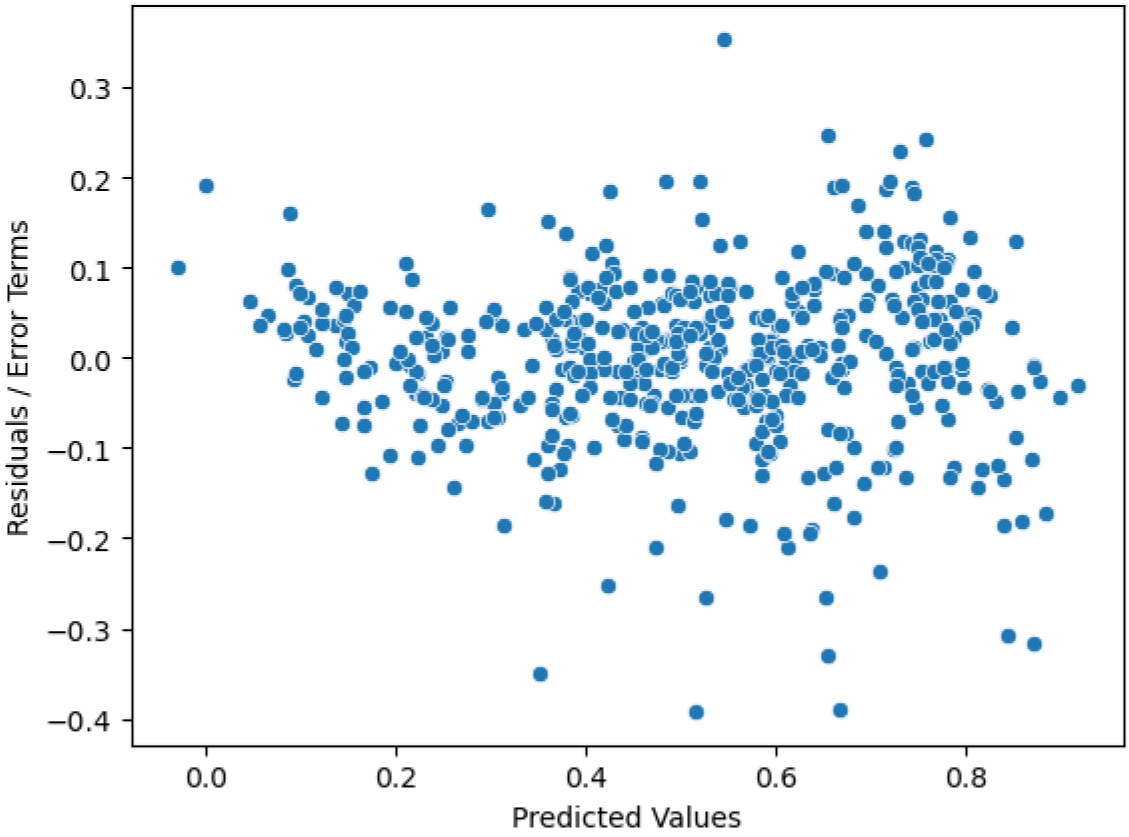


Question 4: How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer: We basically need to validate the below assumptions after building the model on the training set

- There should be a linear relationship between X and y - This has been validated by drawing pairplots & determinining correlation coefficients and using heatmap
- Residuals or error terms should be normally distributed - This has been validated by drawing a histogram of error terms or by plotting a Q-Q (Quantiles - Quantiles) plot
- Verify if the error terms are independent of one another or if there is any visible pattern among the error terms - This has been validated by drawing a scatter plot of residuals against predicted values
- Verify if the error terms have constant variance among them - This has been validated by drawing a scatter plot of residuals against predicted values





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

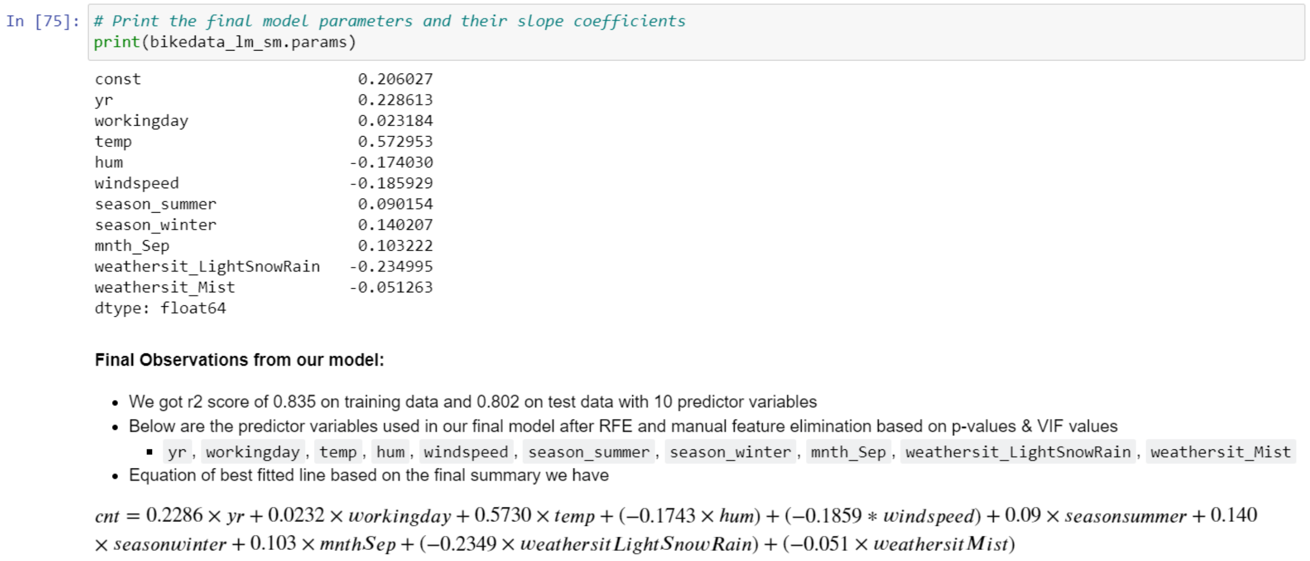
Answer:

- Among the predictor variables, we see that it is `tmp`, `yr` and `windspeed` which have strong correlation coefficients with the target variable
- Below are the correlation coefficients sorted in descending order

```
In [56]: # Observing the correlation coefficients in descending order only with target variable
         train_data_cnt_corr = train_data_corr_current_model.corr()['cnt']
         train_data_cnt_corr.abs().sort_values(ascending=False)
```

```
Out[56]: cnt                          1.000000
         temp                         0.643517
         yr                           0.591508
         windspeed                    0.253794
         weathersit_LightSnowRain     0.226598
         mnth_Sep                     0.201327
         weathersit_Mist              0.175530
         season_summer                0.134343
         workingday                   0.092320
         hum                          0.059993
         season_winter                0.032636
         weekday_Sat                  0.016215
         Name: cnt, dtype: float64
```

- Below are the final model parameters and slope coefficients determined by our model



```
In [75]: # Print the final model parameters and their slope coefficients
         print(bikedata_lm_sm.params)
```

```
const                       0.206027
yr                          0.228613
workingday                  0.023184
temp                        0.572953
hum                        -0.174030
windspeed                  -0.185929
season_summer               0.090154
season_winter               0.140207
mnth_Sep                    0.103222
weathersit_LightSnowRain   -0.234995
weathersit_Mist            -0.051263
dtype: float64
```

Final Observations from our model:

- We got r2 score of 0.835 on training data and 0.802 on test data with 10 predictor variables
- Below are the predictor variables used in our final model after RFE and manual feature elimination based on p-values & VIF values
  - `yr`, `workingday`, `temp`, `hum`, `windspeed`, `season_summer`, `season_winter`, `mnth_Sep`, `weathersit_LightSnowRain`, `weathersit_Mist`
- Equation of best fitted line based on the final summary we have

$$cnt = 0.2286 \times yr + 0.0232 \times workingday + 0.5730 \times temp + (-0.1743 \times hum) + (-0.1859 * windspeed) + 0.09 \times seasonsummer + 0.140 \times seasonwinter + 0.103 \times mnthSep + (-0.2349 \times weathersitLightSnowRain) + (-0.051 \times weathersitMist)$$

General Subjective Questions

Question 1: Explain the linear regression algorithm in detail

Answer:

- Linear Regression is used to predict a numerical target variable using one or more predictor variables by fitting a linear equation. We are assuming that there is a linear relationship between the predictor variable(s) and the target variable
- Our objective in this is to determine the best fit line that minimizes the cost function which is residual sum of sqaures in our scenario. We basically try to ensure that straight line is best fit or close enough to

all points so that the overall sum of residuals is minimum.

  - Below is the linear equation for the same:

    $y = \beta_0 + \beta_1 X_1 + ...+ \beta_n X_n$ Here,

- y - target variable
- $X_1, ... X_n$ are the features or independent variables
- β_0 - Intercept representing the value of y when value of all feature variables is zero
- $\beta_1 ... \beta_n$ - Slopes or coefficients of feature variables. This represents change in y for a unit change in respective feature variable when the values of other feature variables are constant.

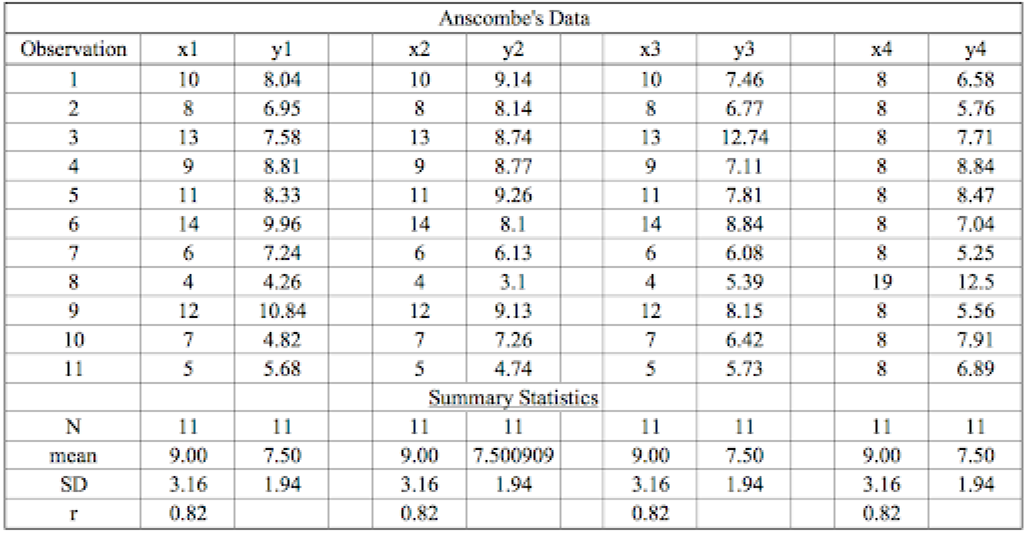
Below are the assumptions of linear regression:

- X and y should have a linear relationship
- Residuals or error terms are normally distributed
- Residuals or error terms are independent and there is a constant variance among all the residuals

Question 2: Explain the Anscombe’s quartet in detail

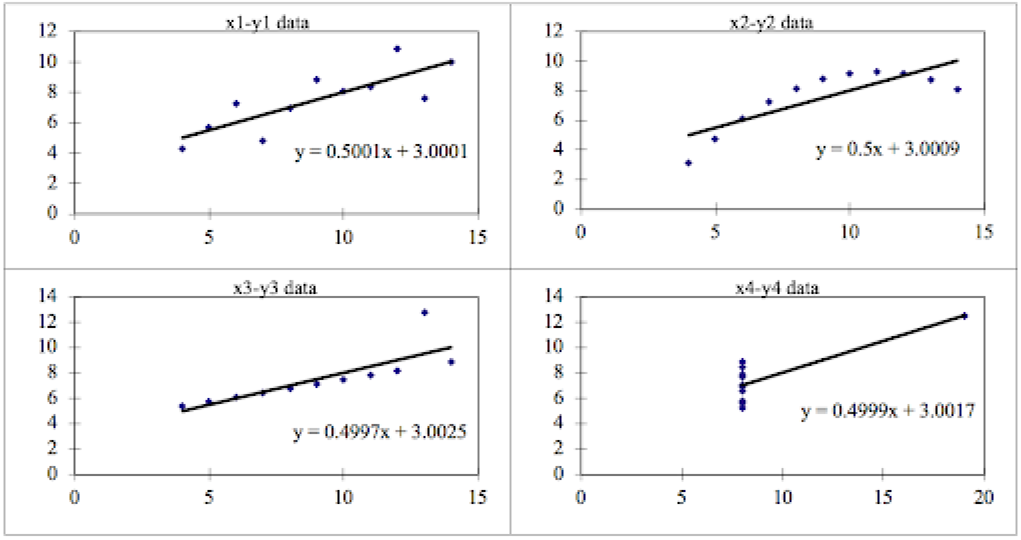
Answer:

- Anscombe's quartet tells us about the importance of visualizing the data before attempting to build any machine learning model on it
- It suggests that the data features must be plotted to see the distribution of the data to identify the anomalies present in the data like outliers, linear relationship etc.
- Some times statistical measures like mean, standard deviation may be same but visualizing the data may help in unearthing the hidden patterns / anomalies present in the data
- In the given example, statistical measures are same across all four datasets. However when we visualize the data, we could see that it is actually differently distributed across the datasets. We can identify linear relationship exists or not, outliers by simply looking at the data visualization



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	



Question 3: What is Pearson's R?

Answer:

- Statistical measure that helps in determining the strength of the linear relationship between two numeric variables.
- Values for this lie in between -1 and 1
  - -1 - This indicates a negative strong relationship. Basically value of y increases when X decresses and it decreases when X increases
  - 0 - No relationship between X and y
  - 1 - Strong correlation between X and y and they both move in the same direction
- So the values close to 0 indicates no linear relationship, where as values close to 1 indicates strong relationship
- This is basically used to quantify the relationship betwen two numeric variables.
- This may not be appropriate when the relationship between the numeric variables is not linear

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

Answer:

- Scaling is a preprocessing technique especially used when we are trying to build a model involving multiple predictor variables or feature variables.

- This is needed to bring all of the predictor variables involved to a common / uniform scale such that the coefficients / slopes can be determined efficiently
- Normalized Scaling
  - In this we will be rescaling the values of given variable to a range between 0 and 1. This will be performed by using minium and maximum values of the given variable as below.

    $X_{normalized} = (X - X_{min})/(X_{max} - X_{min})$
- Standardized Scaling
  - In Standardized scaling, data is transformed in such a way that it has a mean of zero and a standard deviation of 1 by using the below formula

    $X_{standardized} = (X - X_{mean})/X_{std}$
- Below are the key differeneces with respect to normalized scaling and standardized scaling:
  - In normalized scaling, values are rescaled to a fixed range of 0 and 1, whereas standardized scaling transforms the data such that the resultant mean is zero and standard deviation is 1.
  - Standardized scaling is more suited to normally distributed data where as normalized scaling can be used even without knowing how the data is distributed.
  - Standardized scaling preserves the information about distribution and spread of the data even after transforming where as in normalized scaling, we don't see that

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

Answer:

- Variance Inflation Factor (VIF) is typically used to measure the multicollinearity among the predictor variables
- For each independent or predictor variable, we will determine how much variance in it can be explained by using the remaining predictor variables.
- This is calculated using the formula, where R-Squared is the metric obtained by applying linear regression on the remaining predictor variables to predict the given variable.

$VIF = 1/(1 - R^2)$

- If the given variable is highly collinear with other predictor variables, then theoretically VIF might become 1. This happens when the value R-squared approaches 1.

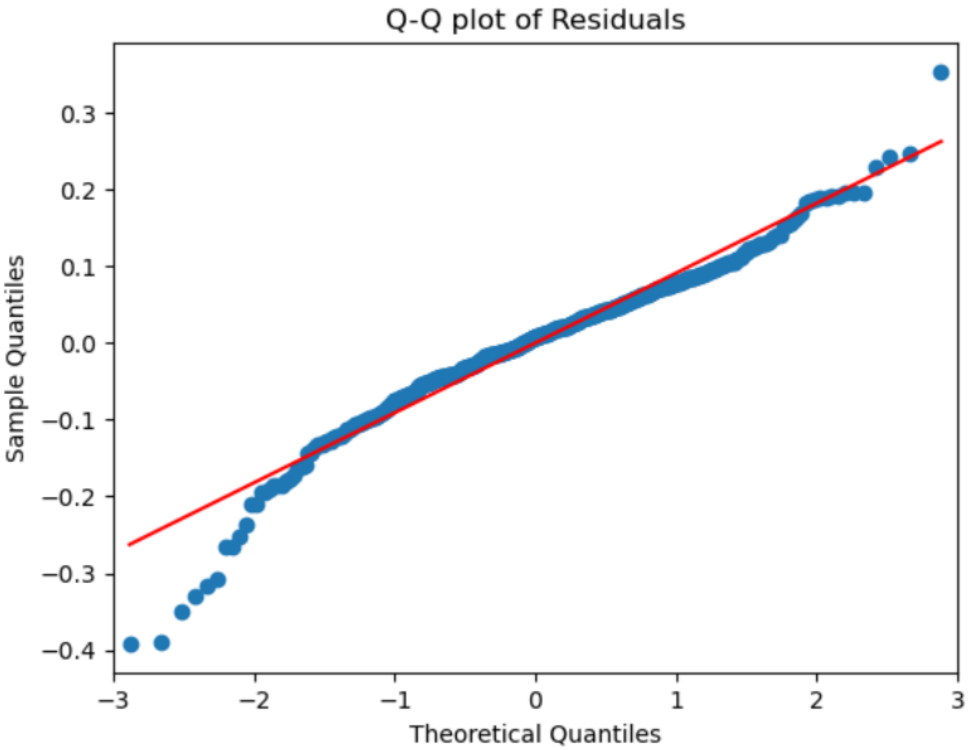
Question 6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression

Answer:

- Q-Q (Quantile-Quantile) plot is a way to assess whether a dataset follows a particular distribution or not. Normally it is used to verify if the given dataset follows a normal distribution or not.
- It basically compares the quantiles in our dataset with the quantiles of theoretical distribution we wanted to compare.
- If the points on the Q-Q plot roughly form a straight line, then it suggests that the data is following that particular distribution.
- Usage in Linear Regression

- In linear regression, this can be used to validate one of the key assumptions in linear regression which is to verify whether the error terms / residuals are normally distributed.
- We can plot a Q-Q plot for the residuals, with the reference line indicating a perfect normal distribution using `line='s'`
- If the points in the Q-Q plot closely follow the reference line, then we can conclude that the residuals are following normal distribution

```
In [65]: sm.qqplot(data=(y_train - y_train_pred_cnt), line='s', ax=plt.gca())
         plt.title('Q-Q plot of Residuals')
         plt.show()
```


Q-Q plot of Residuals
Sample Quantiles
Theoretical Quantiles