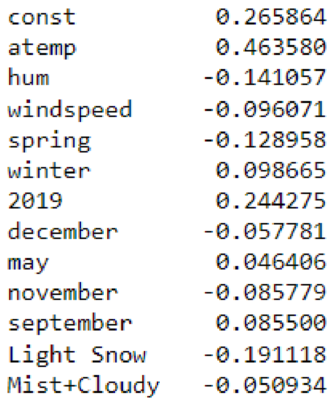
Assignment-based Subjective Questions

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) Final numbers are shown below. Based on the analysis of categorical variables some of the effects are given below:-

- In Seasons, during Spring the count is bound to be less when compared to Winter where the count is on the rise by 0.09
- In months, December and November months sees a drop in count whereas May and September sees a rise
- Weather, Both during Light Snow and Misty days we are seeing a drop in count



```
const          0.265864
atemp          0.463580
hum           -0.141057
windspeed     -0.096071
spring        -0.128958
winter         0.098665
2019           0.244275
december      -0.057781
may            0.046406
november      -0.085779
september      0.085500
Light Snow    -0.191118
Mist+Cloudy   -0.050934
```

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

Ans) When creating dummy variables, we should use drop_first=True to avoid multicollinearity. This normally happens when we are dealing with variables which are having more than one category. Using this will automatically drop the first category.

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

Ans)Temp and atemp has the highest correlation with the target

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

Ans) Linear relationship exists between dependent and non-dependent variables. This can be proven by a scatter plot of x and y. Multicollinearity is another assumption. As per this assumption independent variables should not be correlated – WE use VIF to validate this. VIF is expected to be less than 5. Mean of residual should be 0 is our next assumption and as per our analysis we see that mean of residuals to be-4.2251978816848906e-16 which is more or less 0. Homoscedasticity assumption, plotting the error terms with predicted terms we can check that there should not be any pattern in the error terms. Normality of error is the final assumption we have made and validated using plot which shows that error have a normal distribution

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) The top 3 variables identified are

- 'atemp' - Feeling Temperature: A coeff value of *0.463* indicates that a unit increase in this variables increases the count by *0.463* units
- '2019' - Year: A coeff value of *0.244* indicates that a unit increase in this variable increases the count by *0.244* units
- 'winter' - Season: A coeff value of *0.098* indicates in winter the count is expected to increase by 0.098 units

General Subjective Questions

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

Ans) When we want to predict the outcome of one variable based on its relationship with other variables we go for linear regression. We do this finding a linear equation which fits on the set of data. When there are multiple independent variables the model is extended to :-

$$y = b1x1 + b2x2 + \ldots + bnxn + c$$

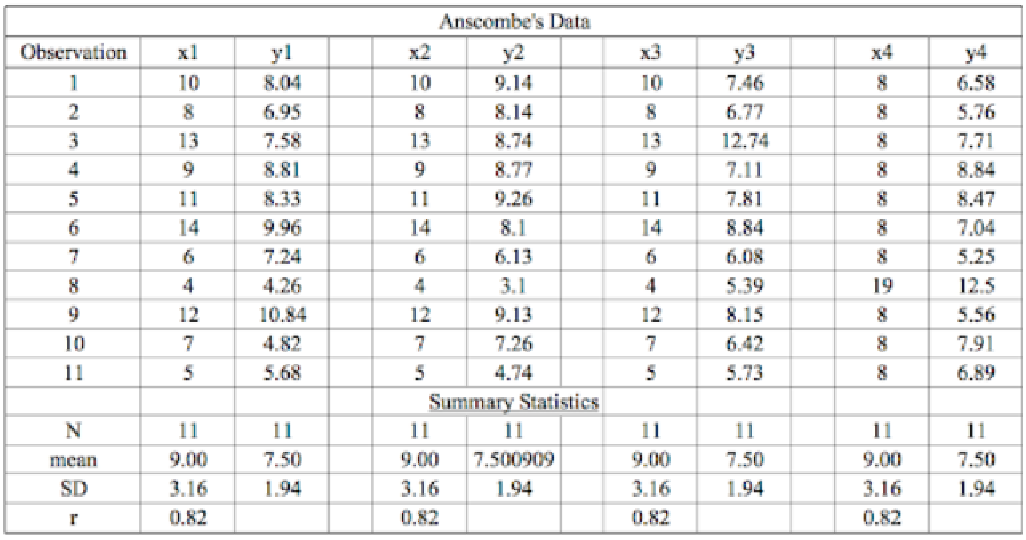
Steps to followed are as follows:

a. Data collection: collect data on both dependent and independent variables
b. Explore: look at the data collected in above steps and complete analysis
c. Preparation: Here we do the clean-up, encode categorical variables and scale variables
d. Training: WE split the overall data as training data and test data. In this stage we use training data to derive the coeffs
e. Evaluation: Here we evaluate the model we have generated using test data
f. Predict: we use trained models and make predictions on new data

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

Ans) Anscombe's quartet are a set of 4 data sets which has similar summary statistics, but shows completely different behaviour when they are plotted as a graph. This emphasizes on the importance on visualising data before applying any algorithms to build models. This also helps us to identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.)

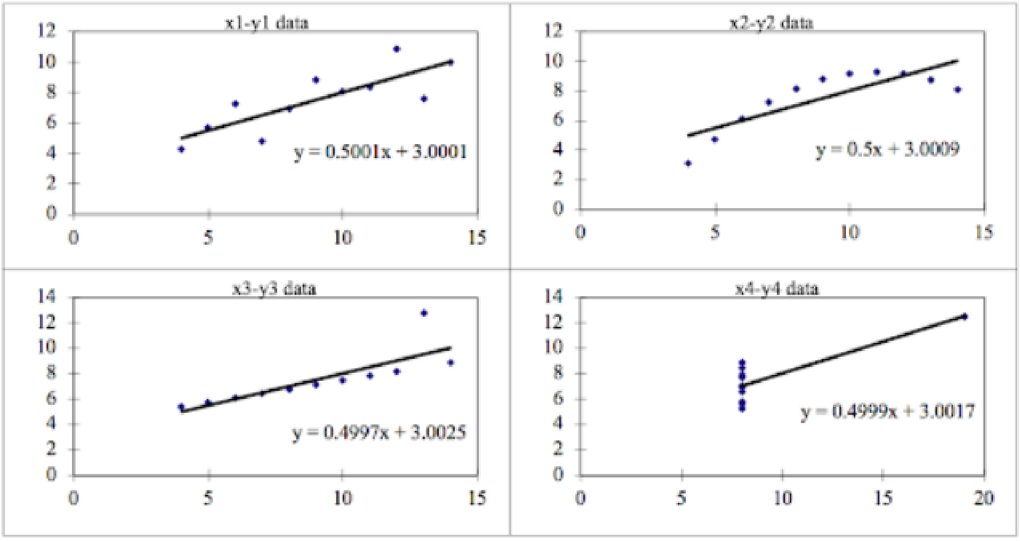
This can be explained using the below dataset. We can see that the summary stats are same.



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	

Now when we plot these, we can see totally different spread.



3. What is Pearson's R? (3 marks)

Ans) Pearson's R is a statistic that measures the linear correlation between 2 sets of data. It gives the information about the magnitude of the association and direction of the relationship. It's a number between -1 and +1, where -1 indicates a negative linear relationship, +1 indicates a positive linear relationship and 0 indicates no linear relationship. Positive correlation indicate that both the variables move in same direction. Negative correlation indicates that as one variable increases the other variable decreases meaning they are inversely related. Commonly used formula to determine the this is given as below:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

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

Ans) During the process of data preparation we go thru a stage called feature scaling. This is a step applied to independent variables to normalize the data within in a particular range. Performing this is important as the collected data set contains features higly varying in magnitude, units etc. IF scaling is not performed then algorithm will take only magnitude into account and not units and hence will result in an incorrect modelling.

Normalized scaling typically means that rescale the values into a range of [0,1]. Standardized scaling typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance). One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

Ans) High VIF indicates high multi-collinearity among the predictor variables. VIF is calculated using the formula

$$\mathrm{VIF} = \frac{1}{1 - R_j^2}$$

Now when the R^2 value comes closer to 1 then the denominator becomes 0 . Dividing a number by 0 is considered as infinity. This normally happens when the jth variable can be perfectly predicted by other variables.

In other words, infinite VIF values occur when there is an exact linear relationship between the predictor variable in question and the other predictor variables in the model.

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

Ans) A Q-Q plot or quantile-quantile plot is a graphical method for comparing 2 probability distribution. A Q-Q plot plots the quantiles of a sample distribution against quantiles of a theoretical distribution. In linear regression Q-Q plot are used for residual analysis – to determine if residuals are normally distributed and if there is a pattern for residuals.

statsmodels.api provide qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively