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: <Your answer for Question 1 goes below this line> (Do not edit)

i) Season: Box plot of cnt for each of the seasons show that cnt is highest in fall, followed by summer, winter and spring.
ii) yr: Box plot shows much higher value for yr=1 (2019) than yr=0
iii) mnth: The monthly variation follows the seasonal pattern, higher during summer and fall months and lower during sprint and winter months
iv) weathersit: It is in the order of 1>2>3>4 where 1 stands for "good-weather/clear-weather", 2 for "average weather", 3 for "bad weather" and 4 for "terrible weather"
v) workingday: There is a slight variation between workingday = 0 and workingday = 1. The 50th percentile and 75th percentile are almost same but the min and max values both are slightly higher for workingday=0.
vi) Weekday: Variation is not significant
vii) Holiday: cnt is higher on non-holidays than on holidays

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)
When we have k categorical levels for a variable, we should use only k-1 dummies to represent them to avoid multicollinearity – so called 'dummy variable trap'. Therefore we use drop_first=True.

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)

Variable 'registered' has highest correlation with cnt.

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)

i. Create a scatter plot of predicted values against the target variable in the test data. This showed a nearly linear relationship.
ii. Variation of error terms: Plot the distribution of error terms using sns.distplot. The variation of error terms of final model showed a near normal distribution with mean almost 0
iii. There is no visible pattern of error terms, which indicates they are independent.
iv. Also checked VIF values for each variable to confirm there is no multicollinearity
v. R-squared value after fitting the model on test data: Close to 1 indicates that the observed values fit closely to fitted values

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)

i. registered
ii. workingday
iii. summer

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)

<Your answer for Question 6 goes here>

A linear regression model makes a prediction by computing a weighted sum of the input features plus a constant bias term.

$Y = \Theta_0 + X_1\ \Theta_1 + X_2\ \Theta_2 + X_3\ \Theta_3 + ... + X_n\ \Theta_n$
Where:
Y is the predicted value
X_1… X_n are the feature values
Θ_1 …. Θ_n are feature parameters (a.k.a weights)

The linear regression algorithms calculate the feature weights so that the model best fits the training set. The performance measure of linear regression model is the root mean square error of the predicted values. The algorithms try to find the feature weights which result in minimum value of the root mean square error. One method is to use a mathematical equation which directly gives the feature weights. This equation is also known as Normal Equation:

$\Theta = (X^T X)^{-1} X^T Y$

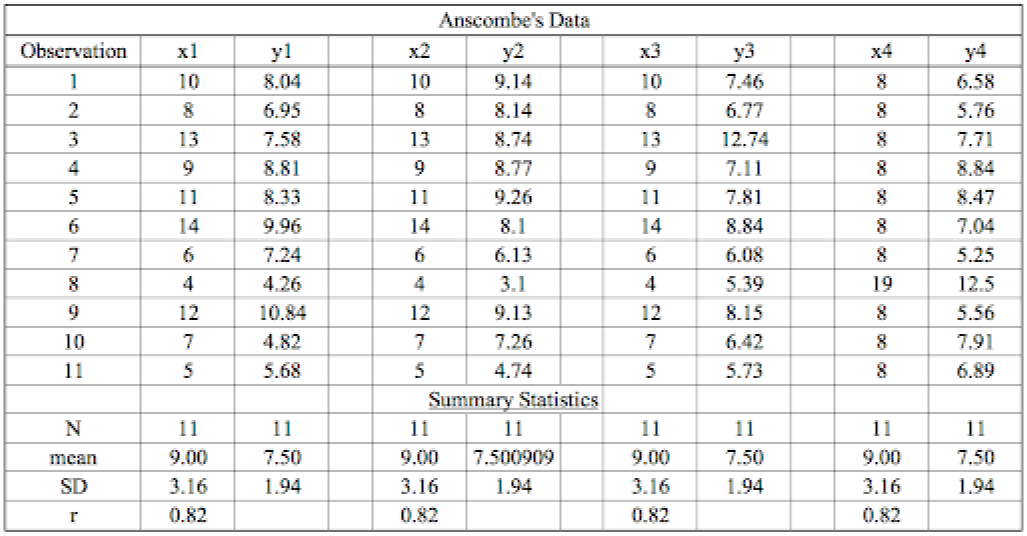
Where Ө is the parameter vector consisting of the feature parameters (aka weights)
X is the feature vector (consisting of the feature values X1… X_n)
Y is the vector of target values

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)

<Your answer for Question 7 goes here>

Anscombe’s quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone. It comprises of a set of four datasets which look different when plotted on scatter plots but having identical values of mean, variance, R-squared and correlation coefficients.

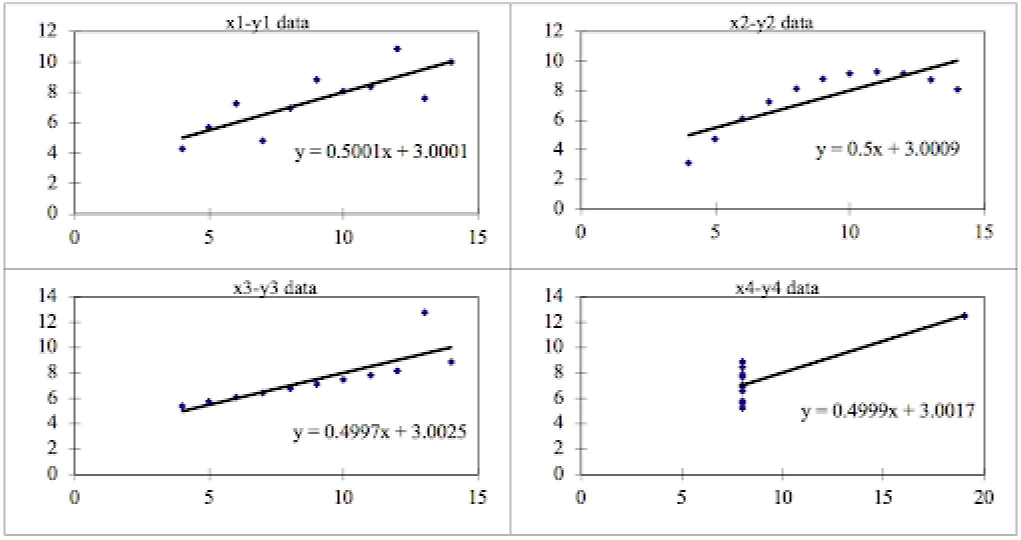
Below image shows the four data sets and their summary values:



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	

Below image shows the plots for the same:



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)

<Your answer for Question 8 goes here>

The standard correlation coefficient is also known as Pearson's R. It determines the strength and direction of a linear relationship between two variables.

It ranges from -1 to +1. When it is close to 1, it means there is a strong positive correlation. When it is close to -1 there is a strong negative correlation. Value of 0 means there is no linear correlation.

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)

<Your answer for Question 9 goes here>

Scaling means getting all parameters in the training data to same scale. In general, machine learning algorithms do not perform well when the input numerical attributes have very different scales. Without scaling it is difficult to compare the weights (coefficients) computed by the linear regression algorithm.

Normalized scaling, also known as Min-max scaling is the simplest: for each attribute, the values are shifted and rescaled so that they end up ranging from 0 to 1.

In standardized scaling, each value is subtracted by the mean value and then divided by the standard deviation. So the standardized values have standard deviation equal to 1. Standardization is not affected by outliers unlike Normalized scaler.

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)

<Your answer for Question 10 goes here>

By definition VIF of i^{th} attribute is equal to $1/(1-R_i^2)$ where R_i^2 is the R-squared value obtained by regressing i^{th} attribute on the remaining attributes.

If VIF is infinite, it is because of $R_i^2 = 0$ which means the ith attribute is completely linearly dependent on others.

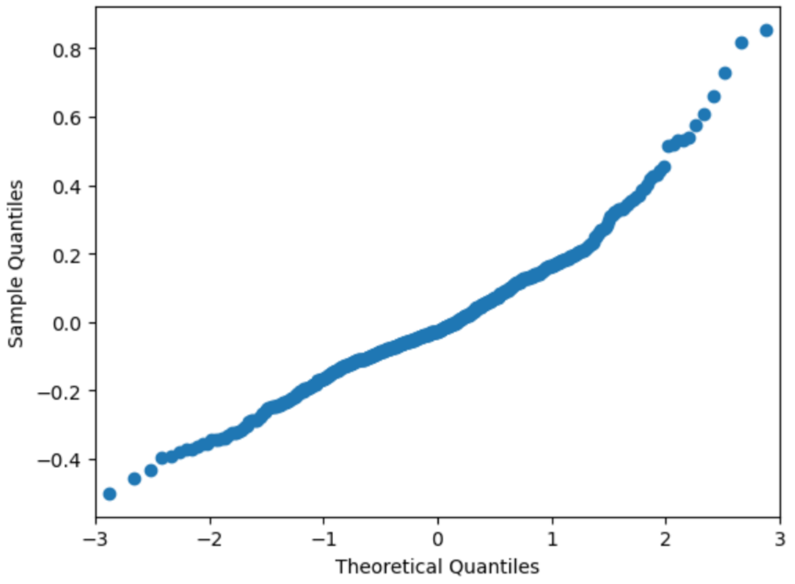
Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>

Quantile-quantile plots (q-q plots) are used to determine to what extent observed data points follow a given distribution. Usually normal distribution is used as the reference. To create a qq plot, we have to plot the quantiles of the observed data points against the quantiles of a normal distribution and see how closely the points fall in a straight line.

After linear regression fit using statsmodel fit API, we can use the qqplot API (from statsmodel) to create a q-q plot of the residuals and check whether they fit a straight line hence whether they follow a normal distribution. For example in the bikesharing assignment, after the final model is arrived at I used the below code and got the plot below:

```
lr = sm.OLS(y_train, X_train_lm.astype(float)).fit()
fig = sm.qqplot(lr.resid)
plt.show()
```


Sample Quantiles
Theoretical Quantiles