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)

- Year on year there is a significant increase of people making use of BikeSharing
- Usage is more if it is not a holiday
- On wokringdays usage is a bit less compared to non-working days
- On weekdays it is almost evenly spread out
- Highest usage is when the weather is clear and the numbers drop as the weather becomes grey
- Summer and Fall have got the highest number of people using the service. During Winter and Spring the numbers start to reduce
- Q1 and Q2 of every year, the footfall is high and gradually decreases as we move into Q3 and Q4 of every year

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

- drop_first=True will help in dropping the redundant column when dummy variables are created. This helps to achieve K-1 columns if K catergories are present in that column

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

- atemp column has the highest correlation with the target variable

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

- All VIF values are less than 5 showing that they are not multi collinear
- Error terms are normally distributed
- There are no patterns displayed in homoscedasticity test
- Variables are linearly related with each other

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

- Year, Temperature and Weather are the top 3 features contributing significantly towards the demand of shared bikes

General Subjective Questions

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

Linear regression is a machine learning algorithm where one of the dependent variable is linearly dependent on the given set of independent variables. This algorithm works by generating a best fit line across the dataset which predicts the output variable value based on the given input values. The hypothesis of linear regression is as follows:

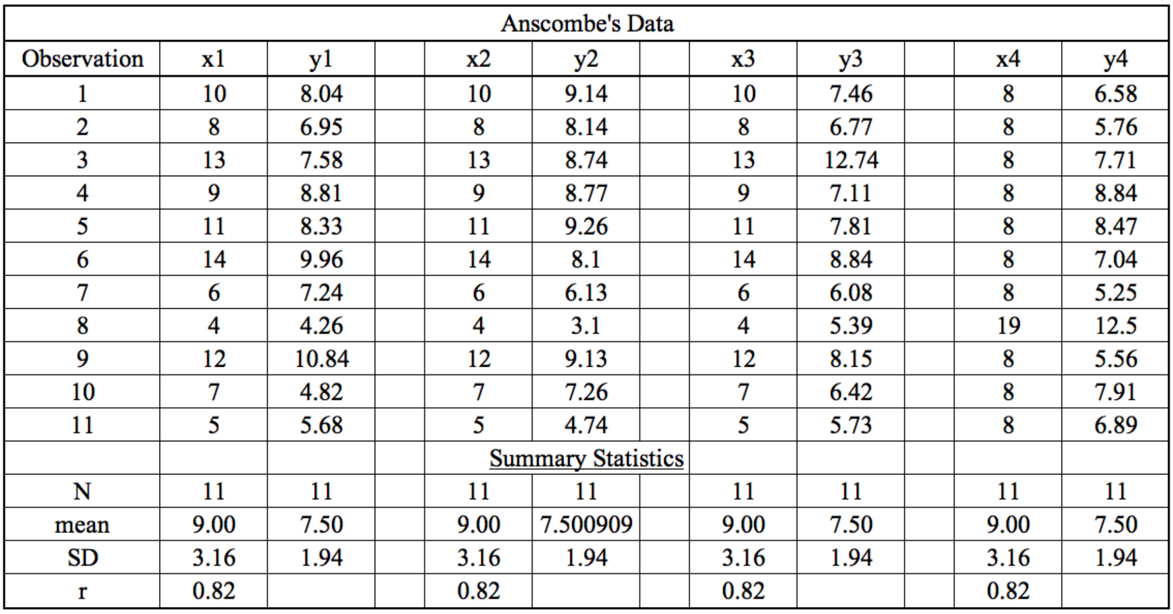
$Y = C + mX$ where,
Y is the output variable
M is the Slope

X is the input variable
C is a constant/Y-intercept

Cost function (J) is used to arrive at the values of C and m. We try to minimise the cost function as much as possible so that the predicted values are almost as close as the actual values. J is also knows as Root Mean Squared Error (RMSE) between the predicted and the actual values. Gradient descent logic is used to minimise the cost function.

J = (1/n) SUMMATION ((predictY - actualY)^2)

2. Explain the Anscombe’s quartet in detail. (3 marks)
Anscombe’s quartet was constructed by Francis Anscombe in 1973. The main purpose of this was to showcase that even with the same statistics across 4 datasets, if plotted in a scatter plot will show completely random distribution. This will in turn affect the linear regression model if built on these datasets. The signifiacnece that Anscombe was trying to show is that it is important to visualise, compare and constrast the data before building a model.



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	

Figure 1: Source - Sparsh Gupta from medium.com

Figure 1 shows the dataset that was considered. The statistics is approximately same for all the 4 datasets

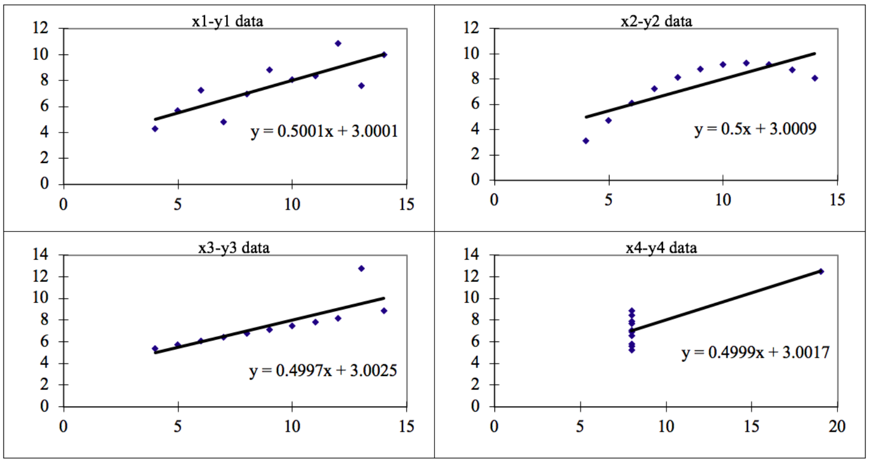


Figure 2: Source - Sparsh Gupta from medium.com

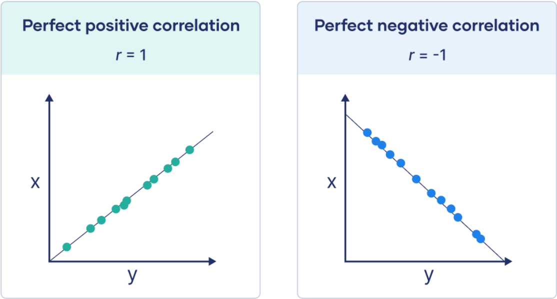
Figure2 shows the same data when plotted in a scatter plot doesn't seem to be consistent.

X1-y1 data – Looks good and a linear regression model can be used in this case
X2-y2 data – Linear regression is not a best fit model for this sort of data
X3-y3 data – Outliers will affect if linear regression model is applied on this
X4-y4 data – Huge set of outliers are present that the linear regression model will not be able to handle

3. What is Pearson's R? (3 marks)
Pearson's R is a way of measuring the linear correlation. It is a value between -1 and +1 measuring the strength and direction of relationship between 2 variables.
The formula to calculate Pearson's coefficient is,

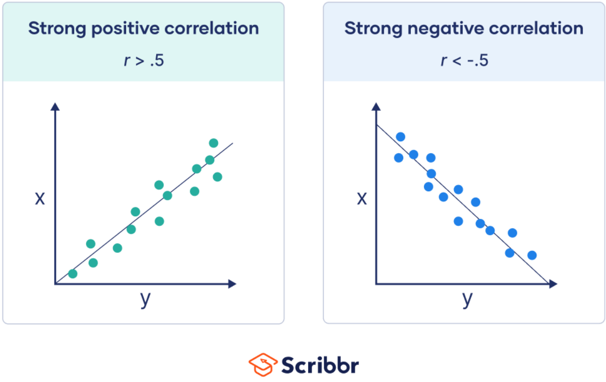
$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Value is either +1 or -1: Strong Positive or Negative Correlation

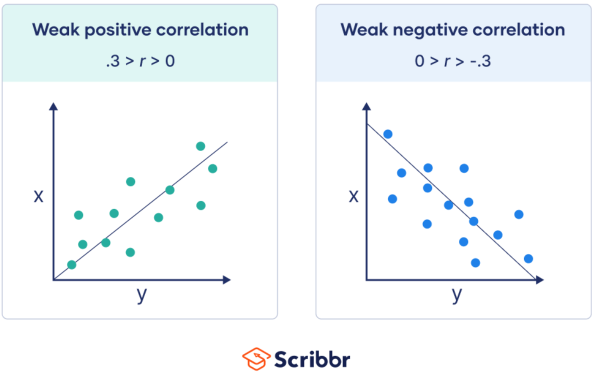


Scribbr

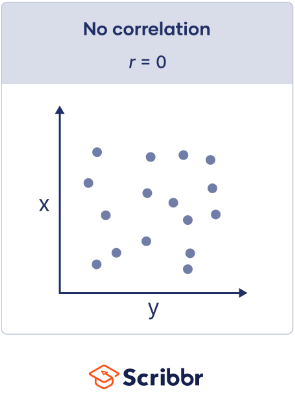
Value is greater than .5 or less than –.5: Significant correlation



Value is between 0 and .3 or between 0 and –.3: Weak correlation



Value is 0: No correlation



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
When working with continuous variables we often tend to see data that are in the 100's scale, in the 10000's scale etc. These when being plot in a graph will be skewed and will be difficult to interpret and will affect the model building. Hence it is advised to bring all continuous variables into a bucket pool of say -1 to +1. This can be achieved by using scaling. Categorical variables doesn't need to be scaled.

S.NO.	**Normalization**	**Standardization**
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scaled values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is often called as Scaling Normalization	It is often called as Z-Score Normalization.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
VIF value of infinity suggests that there is a perfect correlation between the variables. The higher the value of VIF, the higher is the multicollinearity with that variable compared to the other variables. If the VIF value is infinity, then R2 value is 1. This can never happen in real life scenario and it means that our model is over fit. This can be solved by dropping any of the variables having a high p value and that will help in readjusting the VIF value.

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

Quantile-Quantile plot is the process of comparing two probability distributions by comparing their quantiles against each other in a graphical method. There is a 45degree reference line that is plotted and if the two datasets come from a population with same distribution, then the points should lie close to this reference line. As the points start to diverge from the line, it means that they are from populations with different distribution.

While building a linear regression model, it is expected that the datasets considered as train and test are from the same population. This can be validated using Q-Q plot.