Assigned-based Subjective Questions

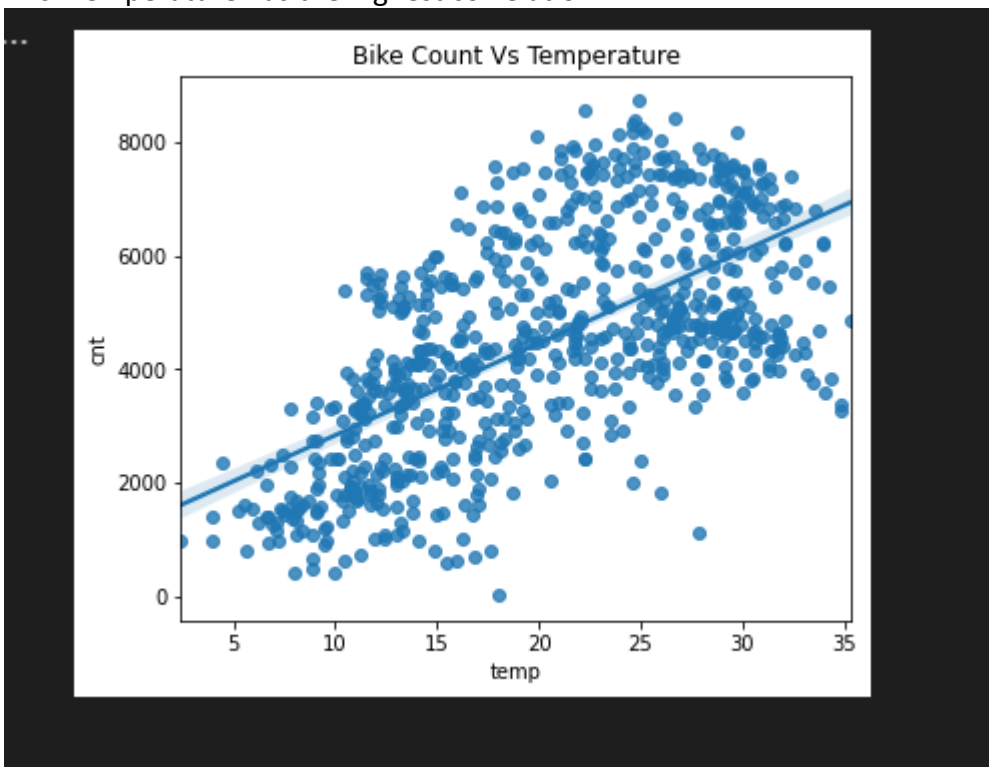
Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans:

1. **Light Rain :** On a light rainy day the bikes have been picked the least, with a slope of -2487
2. **Mist/Cloudy:** On a cloudy day the bike have been picked less, probably the users might assumed that it could rain later in the day
3. **Spring Season:** In Spring season the bike count went down
4. **Winter Season:** Though the bike count didn't spike up much but the bikes were picked up more descently.
5. **Holiday:** If the day was declared a holiday people preferred Netflix and Chill and hence the bike count went down.
6. **Year:** The sales went up from 2018-2019
7. **September:** Compared to other months, the sales went up during this month

Q2. Why is it important to use **drop_first=True** during dummy variable creation?
Ans: Because this will help us drop the extra column during dummy variable creation, as we need n-1 extra columns from n unique values, hence it reduces the correlation created among dummy variables.

Q3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Ans: Temperature has the highest correlation



	cnt	temp	windspeed	hum	atemp
cnt	1.000000	0.627044	-0.235132	-0.098543	0.630685
temp	0.627044	1.000000	-0.158186	0.128565	0.991696
windspeed	-0.235132	-0.158186	1.000000	-0.248506	-0.183876
hum	-0.098543	0.128565	-0.248506	1.000000	0.141512
atemp	0.630685	0.991696	-0.183876	0.141512	1.000000

Q4: How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans:
1. The error state is normally distributed (Plotted a distplot for error terms which was normally distributed)

2. The error state is not correlated with the independent variables/ Error should be independent

3. Homoscedasticity is satisfied (error terms are randomly distributed and they are not dependent)

4. No Multicollinearity (The VIF is less than 5 for all the features)

5. Linearity is there among Target Variable and Independent Variable (Linearity/ Pairplot)

Q5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Ans:

1. Temp: Positive Relationship
2. Weather (Light rain): Negative Relationship
3. Windspeed: Negative Relationship

General Subjective Questions

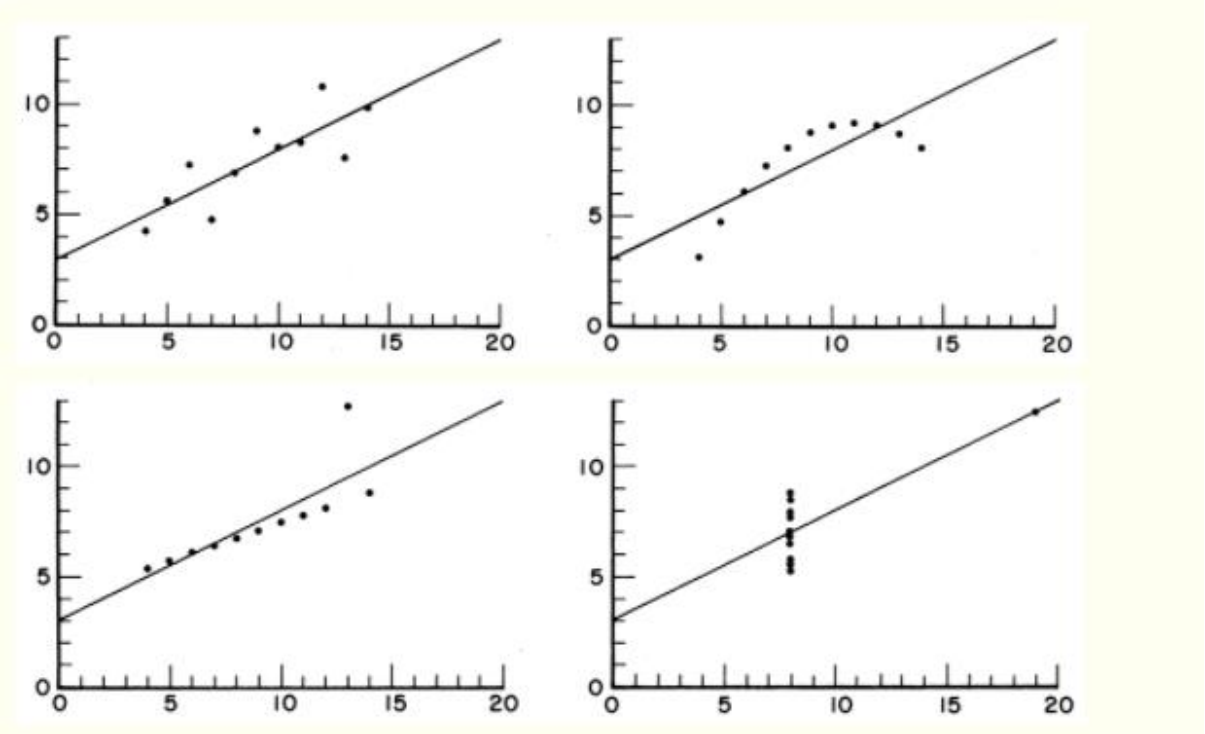
Q1. Explain the linear regression algorithm in detail.
Ans: It is supervised learning algorithm, it performs regression task. In regression task we have a target variable which is also known as dependent variable which needs to be predicted based on dependent variables. Linear Regression algorithm discusses the kind of relationship the two variables posses or have, hence it is called Linear Regression because it measures the relationship between the target variable and independent variables.

While training the model, we will select the line whose cost function is the minimum, the cost function is calculated using the gradient descent algorithm or using Ordinal Least Square Method.

We can measure the accuracy of our hypothesis function by using a **cost function**. This takes an average difference (actually a fancier version of an average) of all the results of the hypothesis with inputs from x's and the actual output y's.
Cost function is represented J(theta0, theta1).
The goal of cost function is to reduce theta0, theta1 also known as intercept and slope.
If we try to think of it in visual terms, our training data set is scattered on the x-y plane. We are trying to make a straight line (y_pred) which passes through these scattered data points.

Our objective is to get the best possible line. The best possible line will be such so that the average squared vertical distances of the scattered points from the line will be the least. Ideally, the line should pass through all the points of our training data set. In such a case, the value of J(theta0, theta1) will be 0. This way we calculate the line which could best possibly relate the independent variable with dependent variable.

Q2: Explain the Anscombe's quartet in detail.
Anscombe's Quarted is defined as a group of four data sets which are identical, however when deep dived and built a regression model of it, can pull us into a dug.
The statistical information for all these four datasets are approximately similar. When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm.
There are four datasets are such that were intentionally created to describe the importance of datavisualition and how many regression algorithm can be fooled at the same time.



Data set		1-3	1	2	3		4	4
Variable		x	y	y	y		x	y
Obs. no.	1 :	10.0	8.04	9.14	7.46	:	8.0	6.58
	2 :	8.0	6.95	8.14	6.77	:	8.0	5.76
	3 :	13.0	7.58	8.74	12.74	:	8.0	7.71
	4 :	9.0	8.81	8.77	7.11	:	8.0	8.84
	5 :	11.0	8.33	9.26	7.81	:	8.0	8.47
	6 :	14.0	9.96	8.10	8.84	:	8.0	7.04
	7 :	6.0	7.24	6.13	6.08	:	8.0	5.25
	8 :	4.0	4.26	3.10	5.39	:	19.0	12.50
	9 :	12.0	10.84	9.13	8.15	:	8.0	5.56
	10 :	7.0	4.82	7.26	6.42	:	8.0	7.91
	11 :	5.0	5.68	4.74	5.73	:	8.0	6.89

Q3: Pearons' R?

Ans: Correlation measures the strength of association between two variables as well as the direction, there are three types of correaltion that are measured.
The Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.
Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

We can calculate Pearson coefficient using the following formulae:

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

Q4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Ans: Scaling is pre processing step which applied on independent variables to normalize the data within a particular range, it also helps in speeding up.

In a real world a dataset might contain data from varying units and range, if scaling is not done then algorithm only takes magnitude in account and not units and hence incorrect modelling

Two types of Scaling:

1. MinMax Scaling/ normalization: (0,1)
2. Standardication (-1,1)

Formula for Standardization: (x-mean(x))/ standardDeviation(x)

Formula for MinMaxScaling: x-min(x)/max(x)-min(x)

Q5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans: This is because there is a perfect correlation, this shows a perfect correlation between two independent variables. In the case of perfect correlation we get R2=1
And formular for VIF is 1/1-R2 which 1/0 also infinity

Q6: Q-Q plot?

It is a scatter plot created by plotting 2 different quantiles against each other, the first quantile ius that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against