Submitted by Achal Kagwad

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)*

Answer: The noted categorical variables which could be of importance in our LR(Linear Regression) Model are: Season, Month and Weathersit. The dependent variable is 'cnt' which is the count of total rental bikes used. By drawing simple box plots against the categories we observed that for season, the fall season saw the maximum usage of bikes with its 25th and 75th percentile ranging between around 4100 and 6100 and the season Spring saw the least usage of bikes with it's 25th and 75th percentile ranging between around 1800 to 3800.

For categorical variable 'month' we see that the usage starts picking up from April and goes till October. Thus these months can be categorized as high usage. Months November, December, Jan and February can be categorized as months with low usage.

Categorical variable Weathersit, A 'clear' weather naturally shows high count/usage with it's 25th Percentile at 3900 and 75th percentile at 6100. When the weather has light snow rain or thunderstorms the count is naturally low with it's 25th and 75th percentile range between 200 to 2000.

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

Answer: The general rule in handling categorical variables in Machine Learning Algorithms is to convert them into dummy numerical variables that are machine understandable. This is done by creating dummy variables(Using one hot encoding and other techniques etc) In General if we have a categorical variable with 'n' levels we would just need (n-1) dummy variables/columns to represent this variable in our model.

Example let's say we have a column named "Furnishing Status' with three categorical levels :"Furnished, Semi Furnished and Unfurnished" We can easily drop column "Furnished" as if the record is not semi furnished and not unfurnished than it's actually furnished. Thus it's important to use drop_first=True during dummy variable creation.

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

Answer: The varibled 'registered' has highest correlation with target variable count, However since 'registered' is multicollinear with 'cnt' we see the next variable as 'temp'(temperature in celsius) having highest correlation of 0.65 with target variable.

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

Answer: The three assumptions that need to be true are with respect to features of error terms, i,e the difference between the actual and predicted values of a given data set. Here in out case: Training Data set. The assumptions are:

- Normal Distribution of Error Terms: We validated this assumption by drawing a histogram of error terms and obtained a normal distribution curve.
- Independence of Error Terms: There should be no pattern observed when a line graph is plotted with error terms with the length of error terms. If there are some patterns observed we conclude that the chosen list of features/variables have some issue such as multicollinearity etc. There was a seasonality pattern in the error plot. This pattern in the errors could probably have been explained by some explanatory variable. We validated this assumption by plotting the necessary graph and actually observing independence of error terms and no pattern observed.
- Constant Variance or homoscadicity: We observed this too when we plotted a graph of error terms with its highest correlated feature which is 'atemp' in our model.

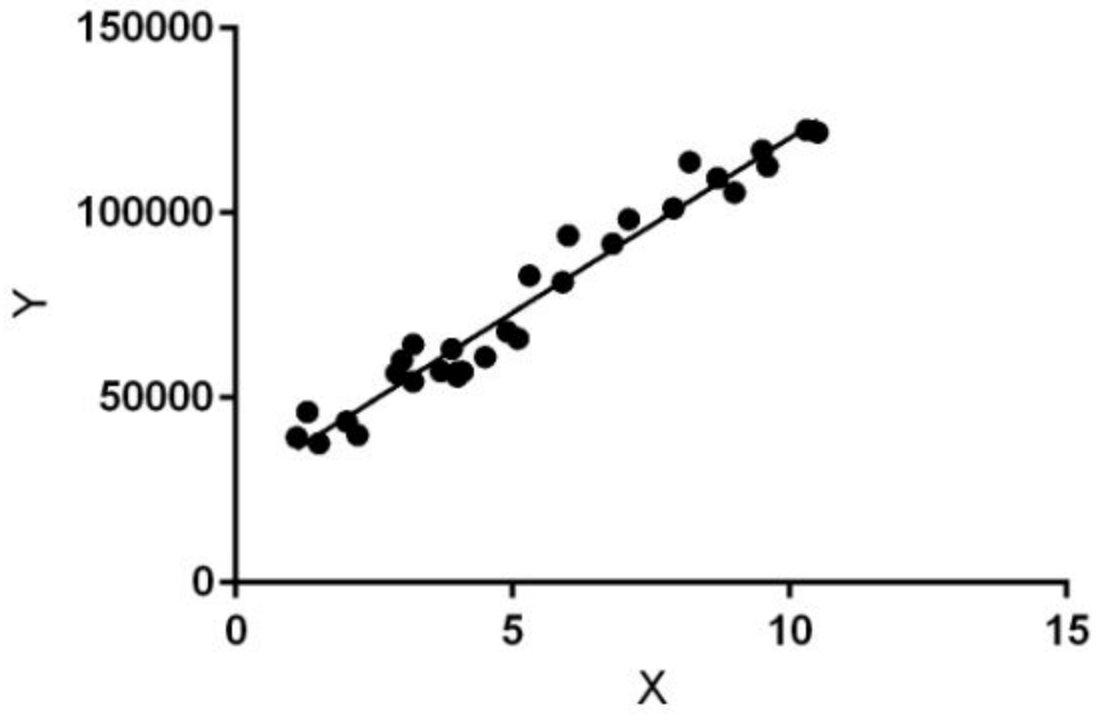
5. *Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)*

Answer: The top three features contributing either positively or negatively towards explaining the demand of shared bikes are 1) atemp 2)yr 3)spring season. Atemp and yr contribute positively meaning increase in their units correspond to increase in demand of shared bikes and spring season contributes negatively meaning increase in it's unit corresponds to decrease of demand of shared bikes. This observation should have some business value in the real world.
The actual equation of Linear Regression which we obtained by the model is given by :
cnt = 0.241153 \times yr + 0.450626 \times atemp + (-0.140749) \times spring + 0.194207

General Subjective Questions:

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

Answer: **Linear Regression** is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.
In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

Hypothesis Function for Linear Regression is given by $y = \beta(0) + \beta(1)x$

While training the model we are given :

x: input training data (univariate – one input variable(parameter))

y: labels to data (supervised learning)

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best $\beta 1$ and $\beta 2$ values.

$\beta 1$: intercept

$\beta 2$: coefficient of x

Once we find the best $\beta 1$ and $\beta 2$ values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

How to update θ1 and θ2 values to get the best fit line ?

Cost Function (J):

By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the θ1 and θ2 values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).

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

Answer: Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

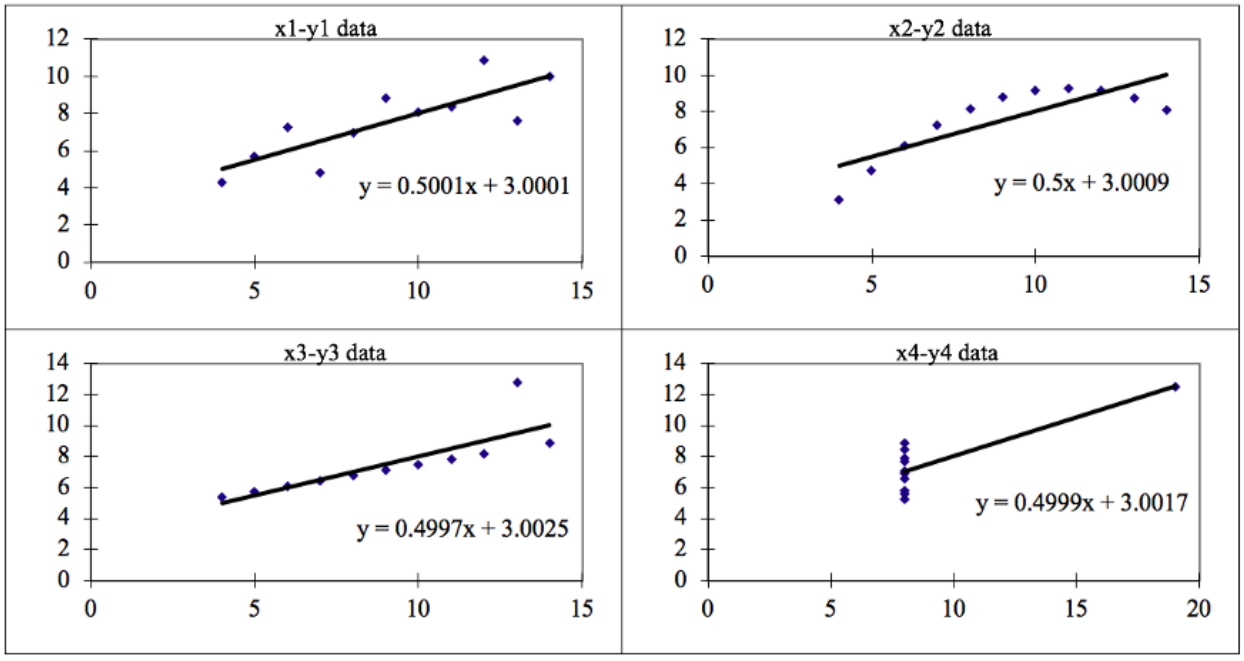
This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets. These four plots can be defined as follows:

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

The statistical information for all these four datasets are approximately similar and can be computed as follows:

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	

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 which is fooled by these peculiarities and can be seen as follows:



Explanation of this above output:

- In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.

- In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

Conclusion:We have described the four datasets that were intentionally created to describe the importance of data visualisation and how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.

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

Answer: Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.
Questions Answered:
Do test scores and hours spent studying have a statistically significant relationship>
Is there a statistical association between IQ scores and depression?
Assumptions:

1. Independent of case: Cases should be independent to each other.
2. Linear relationship: Two variables should be linearly related to each other. This can be assessed with a scatterplot: plot the value of variables on a scatter diagram, and check if the plot yields a relatively straight line.
3. Homoscedasticity: the residuals scatterplot should be roughly rectangular-shaped.

Properties:

1. Limit: Coefficient values can range from +1 to -1, where +1 indicates a perfect positive relationship, -1 indicates a perfect negative relationship, and a 0 indicates no relationship exists..
2. Pure number: It is independent of the unit of measurement. For example, if one variable's unit of measurement is in inches and the second variable is in quintals, even then, Pearson's correlation coefficient value does not change.
3. Symmetric: Correlation of the coefficient between two variables is symmetric. This means between X and Y or Y and X, the coefficient value of will remain the same.

Degree of correlation:

1. Perfect: If the value is near ± 1, then it said to be a perfect correlation: as one variable increases, the other variable tends to also increase (if positive) or decrease (if negative).
2. High degree: If the coefficient value lies between ± 0.50 and ± 1, then it is said to be a strong correlation.
3. Moderate degree: If the value lies between ± 0.30 and ± 0.49, then it is said to be a medium correlation.
4. Low degree: When the value lies below + .29, then it is said to be a small correlation.
5. No correlation: When the value is zero.

Q4. *What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?(3 marks)*

Answer: Scaling of variables is an important step because, as you may have noticed, the variable 'area' is on a different scale with respect to all other numerical variables, which take very small values. Also, the categorical variables that you encoded earlier take either 0 or 1 as their values. Hence, it is important to have everything on the same scale for the model to be easily interpretable.

Scaling of variables is an important step in data preparation because while building our Linear Regression Model, we have some variables/features ranging with high weight or value, whereas some numerical variables have very small values such as ranging between say 1 to 10. Also there would be some binary categorical variables which are encoding into 1s and 0s. So it is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation. So it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale.

As you know, there are two common ways of rescaling:

1. Min-Max scaling
2. Standardisation (mean-0, sigma-1)

Difference between Min Max and Standardisation Scaling/What to use when?:

The formulas given are:
Standardization:x= x-mean(x)/sd(x). Here Mean=0 and sd=1. Standardization transforms your data such that the resulting distribution has a mean of 0 and a standard deviation of 1
Min Max Scaling(Normalization):x=x-min(x)/max(x)-min(x). Here all values come in the range of {0 to 1}. That is min=0 and max=1
As we can see from the formulas: the advantage of Standardisation over the other is that it doesn't compress the data between a particular range as in Min-Max scaling. This is useful, especially if there are extreme data points (outliers).

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

Answer: When the value of VIF is 'inf', it means that the variable is perfectly correlated with some other variable in the machine learning model. In VIF, each feature is regression against all other features. If R2 is more which means this feature is correlated with other features. [0]

- VIF = 1 / (1 – R2)
- When R2 reaches 1, VIF reaches infinity

- We try to remove features for which VIF > 5; If VIF=5, it denotes R2 value of 0.8.

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

Answer: Statistics, Q-Q(quantile-quantile) plots play a very vital role to graphically analyze and compare two probability distributions by plotting their quantiles against each other. If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x.

The most fundamental question answered by QQ plot is: Is this curve normally distributed?

You can tell the type of distribution using the power of the Q-Q plot just by looking at the plot. In general, we are talking about Normal distributions only because we have a very beautiful concept of **68–95–99.7 rule** which perfectly fits into the normal distribution So we know how much of the data lies in the range of first standard deviation, second standard deviation and third standard deviation from the mean. So knowing if a distribution is Normal opens up new doors for us to experiment with the data easily. Secondly, Normal Distributions occur very frequently in most of the natural events which have a vast scope.

In Linear regression one of the assumptions that we have to prove is normal distribution of error terms. In Linear Regression We can use QQ plot to plot residuals(error terms) and check if the residual data points fall on a straight line, if they do the residuals are normally distributed.