Linear Regression Subjective Questions and Answer

Assignment-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: In the final regression model, there were many categorical variables of different catagories, weather parameters, like, Temperature, Humidity or windspeed plays major role where as seasons and months also plays vita roal in business. So looking at the table below, we can conclude that these variables are significant and plays important role

```
                         OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:             0.843
Model:                            OLS   Adj. R-squared:        0.838
Method:                 Least Squares   F-statistic:           165.8
Date:                Sat, 09 Dec 2023   Prob (F-statistic):   1.45e-186
Time:                        19:14:43   Log-Likelihood:      511.12
No. Observations:                 510   AIC:                  -988.2
Df Residuals:                     493   BIC:                  -916.3
Df Model:                          16
Covariance Type:            nonrobust
==============================================================================
        coef     std err      t      P>|t|     [0.025     0.975]
------------------------------------------------------------------------------
const   0.2968     0.034     8.663     0.000     0.229     0.364
yr      0.2322     0.008    28.453     0.000     0.216     0.248
temp    0.4389     0.033    13.174     0.000     0.373     0.504
hum    -0.1598     0.038    -4.186     0.000    -0.235    -0.085
wspd   -0.1856     0.026    -7.149     0.000    -0.237    -0.135
spring -0.0723     0.019    -3.794     0.000    -0.110    -0.035
winter  0.0924     0.018     5.122     0.000     0.057     0.128
w_2    -0.0559     0.011    -5.308     0.000    -0.077    -0.035
w_3    -0.2460     0.027    -9.262     0.000    -0.298    -0.194
m_10    0.0507     0.018     2.881     0.004     0.016     0.085
m_3     0.0660     0.015     4.385     0.000     0.036     0.096
```

m_4	0.0705	0.020	3.498	0.001	0.031	0.110
m_5	0.1011	0.019	5.410	0.000	0.064	0.138
m_6	0.0605	0.019	3.143	0.002	0.023	0.098
m_8	0.0560	0.018	3.096	0.002	0.020	0.092
m_9	0.1199	0.018	6.840	0.000	0.085	0.154
Sunday	0.0204	0.011	1.839	0.067	-0.001	0.042

==

Omnibus:	92.136	Durbin-Watson:	1.952
Prob(Omnibus):	0.000	Jarque-Bera (JB):	251.023
Skew:	-0.883	Prob(JB):	3.10e-55
Kurtosis:	5.949	Cond. No.	19.6

==

Q2) Why is it important to use drop_first=True during dummy variable creation?

Ans: When creating dummy variables, “drop_first=True” need to used so to avoid the "dummy variable trap" or "dummy variable multicollinearity."

The dummy variable trap occurs when dummy variables created for categorical featu res are highly correlated, meaning one can be predicted from the others.

This multicollinearity can cause issues in statistical analyses, particularly in linear reg ression.

By setting “drop_first=True”, you eliminate one of the dummy variables, and this help s to

avoid multicollinearity. In general, it is a good practice to “drop_first=True” so to get in terpretability of the model, especially when using linear regression or other models th at are

sensitive to multicollinearity.

Q3) Looking at the pair-plot among the numerical variables, which one has the highest cor relation with

the target variable?

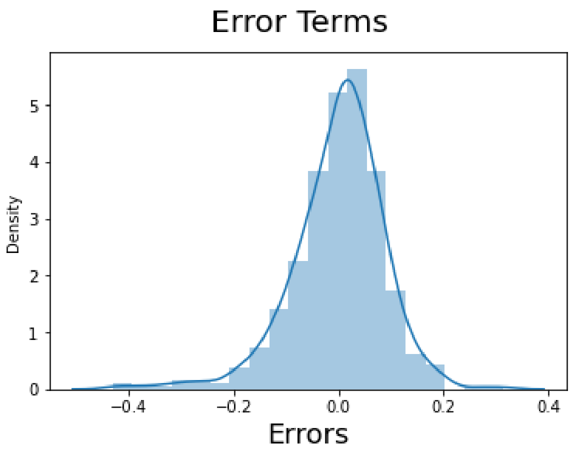
Ans : Numerical Variable’s pair plots clearely that variable ‘tmp’ and ‘atmp’ are in goo d coorelation with count.

Looking at correlation heatmap it is clearely seen that tmp and atmp are well correlat ed with count, with correlation values 0.627 and 0.671 respectively

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

Ans: To validate assumptions of Linear Regression model, there are different method s few of them are verified here in our experiment.

1)Residual Analysis: We need to check if the error terms are also normally Distributed, which is in fact, one of the major assumptions of linear regression Here is the histogram plot of the error terms and this is what it looks like



The Reciduals follows Normal distrubration with mean =0

2) Linear relationship between predictor variables and target variable:
This is happening because all the predictor variables are statistically significan t (p-values are less than 0.05). Also, R-Squared value on training set is 0.827 and adjusted R-Squared value on training set is 0.817 This means that variance in data is being explained by all these predictor variables.

3)Errors are Independent of each other
The predictor variables are independent of each other and Multicollinearity iss ue is not there because VIF for all predictor variables are below 5.

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

Ans: Top 3 features significantly contributing towards demand of shared bikes are:

	Parameters	Coef
1)	Temp	0.4393
2)	year	0.2317
3)	winter	0.0964

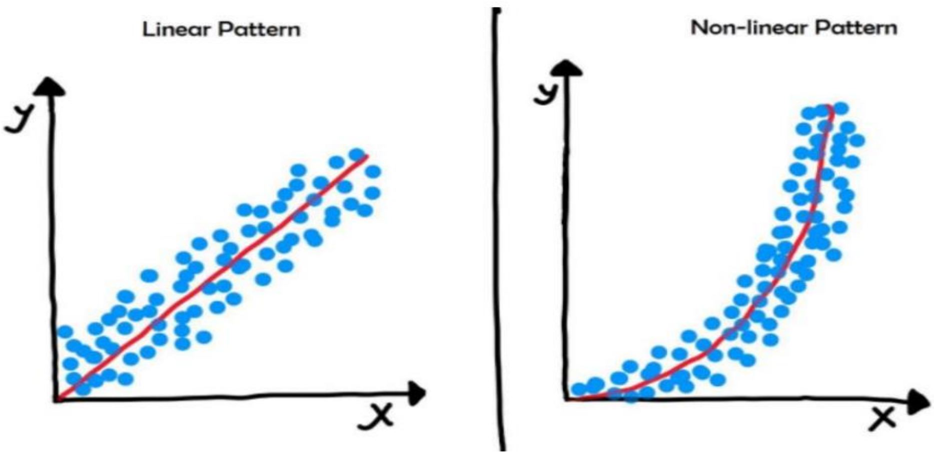
General Subjective Questions

Q1). Explain the linear regression algorithm in detail.

Ans: Linear Regression finds the best linear relationship between the independent an d dependent variables. It is a method of finding the best straight-line fitting to the gi ven data. In technical terms, linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dep endent variables. It is mostly done by the Sum of Squared Residuals Method. T he assumptions of linear regression are:

a. The assumption about the form of the model:

It is assumed that there is a linear relationship between the dependent and ind ependent variables.

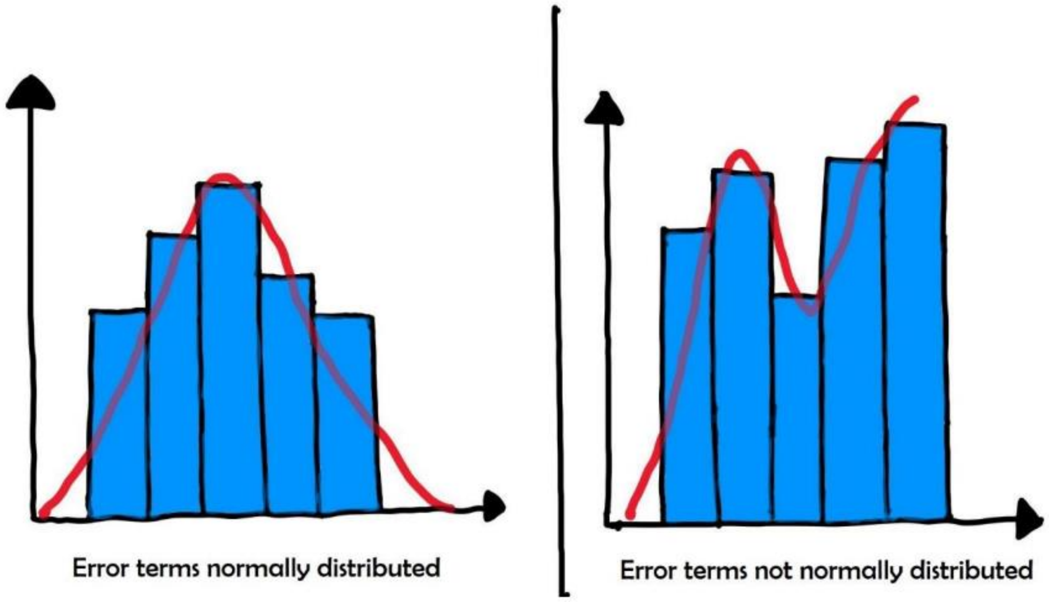


b. Assumptions about the residuals:

1) Normality assumption: It is assumed that the error terms, are normally distributed.2) Zero mean assumption: It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.

3)Constant variance assumption: It is assumed that the residual terms have the same (but unknown) variance, sigma square. This assumption is also known as the a ssumption of homogeneity or homoscedasticity.

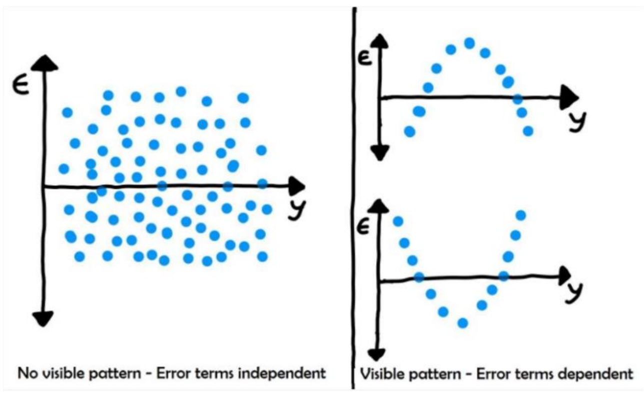
4)Independent error assumption: It is assumed that the residual terms are independe nt of each other, i.e., their pair-wise covariance is zero.

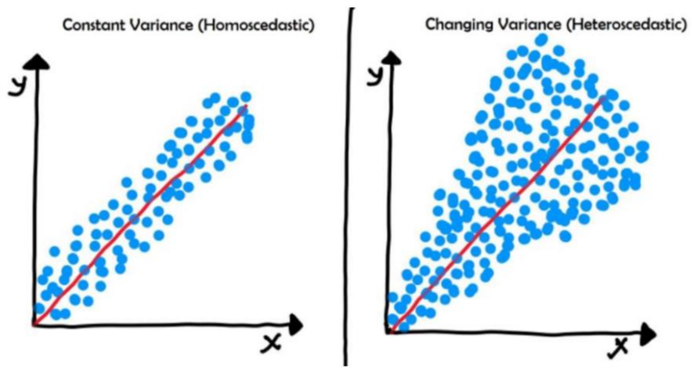


c. Assumptions about the estimators:

1) The independent variables are measured without error.

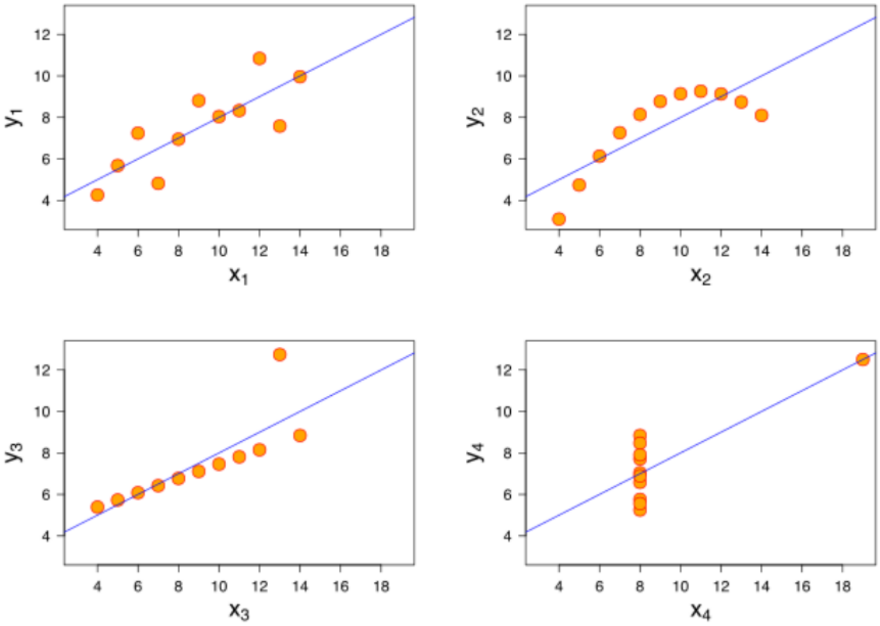
2) The independent variables are linearly independent of each other, i.e., there is no multicollinearity in the data.





Q2) Explain the Anscombe’s quartet in detail. (3 marks)

Ans: Anscombe's quartet comprises four data sets that have nearly identical simpl e descriptive statistics, yet have very different distributions and appear very differe nt when graphed



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed.

1) The first scatter plot (top left) appears to be a simple linear relationship, correspon ding to two variables correlated where y could be modelled as gaussian with mean lin early dependent on x.

2) The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficie nt of determination would be more appropriate.

4)In the third graph (bottom left), the distribution is linear, but should have a different r egression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.8164) the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

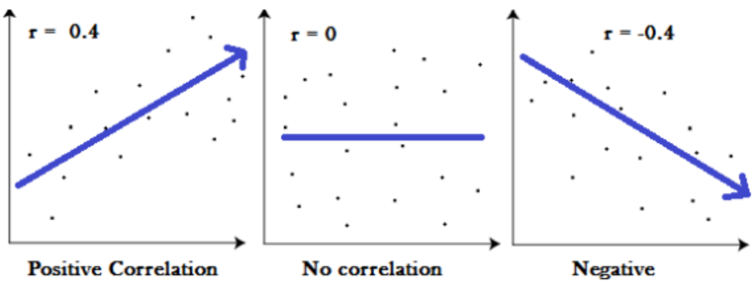
Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Q3) What is Pearson’s R?

Ans: Pearson’s R or correlation coefficient is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation.



1) A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.

2) A correlation coefficient of -1 means that for every positive increase in one variabl e, there is a negative decrease of a fixed proportion in the other. For example, the am ount of gas in a tank decrease in (almost) perfect correlation with speed.

3) Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

Q4) What is scaling? Why is scaling performed? What is the difference between normalize d scaling and standardized scaling?

Ans: Scaling is a method used to normalize the range of independent variables or features of data.

The range of values of raw data varies widely, in some machine learning algorithms, Objective functions will not work properly without normalization for example, many classifiers calculate the distance between two points by the Euclidean distance.

If one of the features has a broad range of values, the distance will be governed by this particular feature.

Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.

Another reason why feature scaling is applied is that gradient descent converges mu ch faster with feature scaling than without it.

Normalization: Also known as min-max scaling or min-max normalization, it is the sim plest method and consists of rescaling the range of features to scale the range in [0, 1]. The general formula for normalization is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Here, max(x) and min(x) are the maximum and the minimum values of the feature respectively

Standardization: Feature standardization makes the values of each feature in the dat a have zero mean and unit variance. The general method of calculation is to determin e the distribution mean and standard deviation for each feature and calculate the new data point by the following formula:

$$x' = \frac{x - \bar{x}}{\sigma}$$

Here, σ is the standard deviation of the feature vector, and x̄ is the average of the fea ture vector.

Q5) You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: If there is perfect correlation, then VIF = infinity.

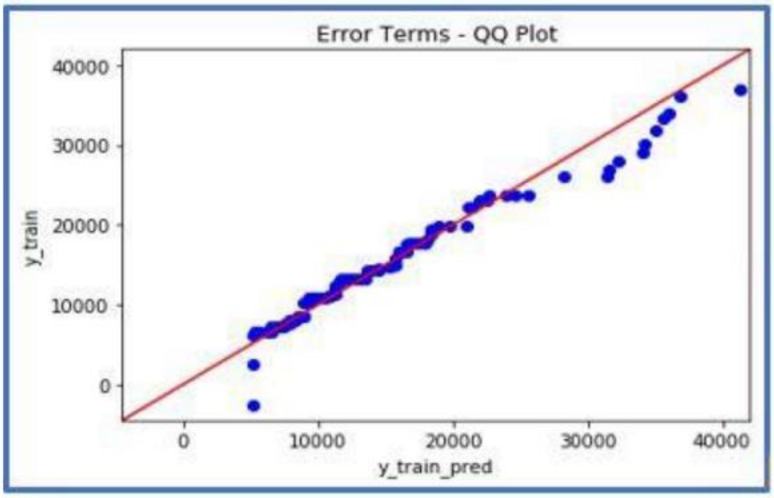
This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.
To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables including VIF

Q6) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression

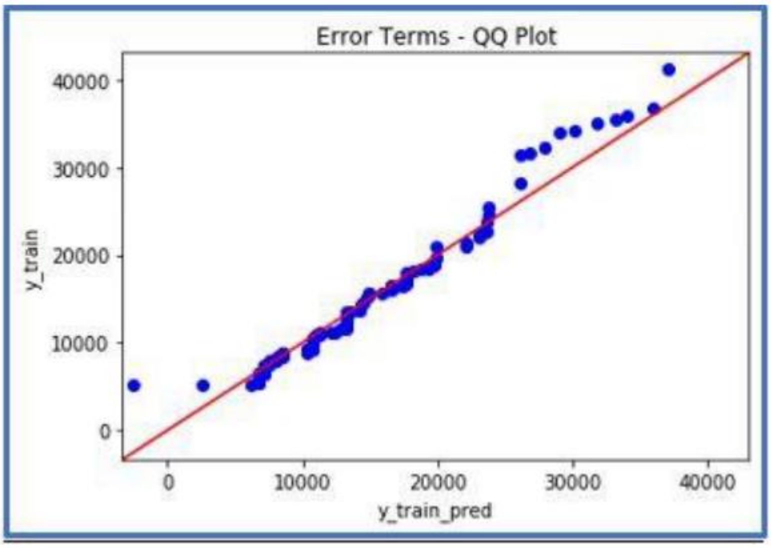
Ans: Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile.
For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.
The purpose of Q-Q plots is to find out if two sets of data come from the same distribution.
A 45-degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

This helps in a scenario of linear regression when we have training and test data set received
separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Interpretation: A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Below are the possible interpretations for two data sets.
1) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
2) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



3) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



4) Different distribution: If all point of quantiles lies away from the straight line at an a ngle of 45 degree from x -axis stats models.

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

- Referances

https://www.geeksforgeeks.org/ml-linear-regression/
https://www.geeksforgeeks.org/anscombes-quartet/
https://www.scribbr.com/statistics/pearson-correlation-coefficient/
https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35
https://www.itl.nist.gov/div898/handbook/eda/section3/qqplot.htm