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?
   Ans:
   1 – The bike share count has increased in the year 2019.
   2 – The business is more in the month from May to Oct in both the years
   3 – Lowest business is in season spring.
   4 – There is negligible business on holiday.
   5 – Business is more on working day.

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

   Ans:

   If we don't use drop_first then n dummy variables will be created, and these n dummy predictors are themselves correlated which leads to multicollinearity. Therefore, we should create n-1 dummy variable for a categorical variable having n distinct values. Dummy variables will have only 0 and 1 values. If all n-1 dummy variables have 0 value indicates that last dummy variable will have value 1. So, avoid multicollinearity we need to use drop_first to delete one dummy variable.

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

   Ans: looking at pair-plot, **temp** has the highest correlation with target variable.

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

   Ans:

   1 – After plotting the histogram of error terms it is observed that error terms are normally distributed at mean = 0.

   2 – The **VIF values of independent variables except temp are less than or close to 5**. So we can say all predictors are not correlated with each other to form multicollinearity. Predictor **`temp`** is most linearly correlated with output variable and also important feature for business evaluation, therefore we have kept it in the model. After dropping **`temp`** predictor there is significant drop in the F-Statistic & R^2 measures.

   3 – After plotting the scatter plot between continuous independent variables and error terms we can conclude that are is no specific pattern followed by the error terms. So, we can say that error terms are independent of each other.

   4 – Since all the coefficients are non-zero, x and y have linear relationship.

   5 – When we checked the error term variance, it is found that it is constant between -0.3 to +0.3 for both `temp` and ` windspeed ` continuous features.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   Ans:
   Below is the list of features affecting (positively and negatively both) significantly towards target variable (demand of shared bikes).
   Top 3: temp(+), light_snow(-), yr(+)
   Except top 3: windspeed(-), Mist_Cloudy(-), sep(+), spring(-), winter(+), Sunday(+), holiday(-), workingday(+), july(-), jan(-), summer(+), oct(+)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Ans:

   In regression we predict the continuous variable.

   The assumptions of linear regression are as follows

   1 – X and Y should have linear relationship

   2 – All independent variables should not have correlation with each other.

   3 – error terms should be normally distributed with mean as 0.

   4 – error terms are independent of each other, i.e., next error is not dependent on previous error.

   5 – error terms should have constant variance.

   Linear regression is used for only interpolation (maximum range of X values we tested) not extrapolation.

   Equation of straight line = Y = beta0 + beta1 * x (beta0 = intercept, beta1 = slop)

   Equation of linear regression Y = c + m1*x1 + m2x2 +……. + mn*xn

   Cost function is used to identify the relation between the output and predictor variables. Cost function is mathematical function which is to be minimized to get the optimal value of beta0 and beta1.

   P value also known as predictive power helps to determine that the relation, we observe in the sample also exists in larger population. P value for each independent variable tests the null hypothesis (i.e., variable has no correlation with the dependant variable). P value should be approximate zero i.e., <= 0.05 for better correlation. P value shows result is statistically significant or not.

   We can predict the model using R^2

   R^2 = 1 – (RSS/TSS)

   RSS = sum of residual squares. Means (y actual - y predicted) ^2 for all y

   TSS = total sum of squares. Means (y actual - y mean) ^2 for all y

   ESS (explained sum of squares) = (y pred – y mean) ^2

   We can also predict the model using RSE (Residual Square Error)

   RSE = sqrt (RSS/df)

   df is sigma i.e., n-2 where n is no of data points

2. Explain the Anscombe's quartet in detail.

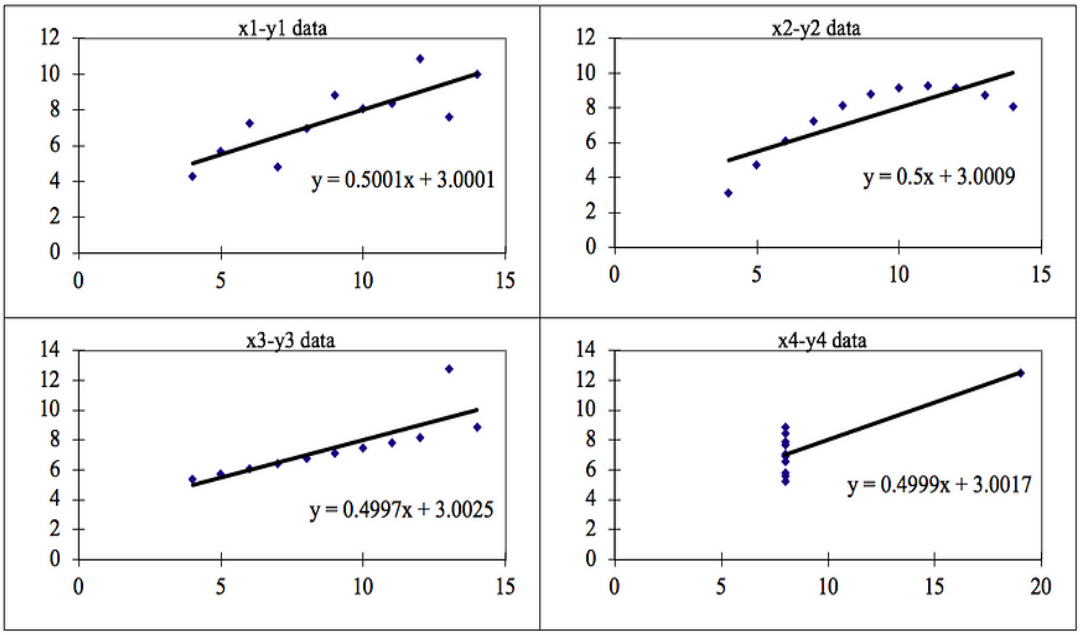
Ans:

Anscombe quartet can be defined as a group of four datasets which are nearly identical in simple descriptive statistics. But there are some particularities in the dataset that fools the regression model if built. They appear very differently when potted on scatter plots.

This tells us the importance of visualizing the data before applying various algorithms to build models. **It suggests that 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**.

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 scatter plot, all datasets generate a different kind of plot that is not interpretable by any regression algorithm which is fooled by these particularities.



Regression module can be fooled by intentionally created data sets. 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?

Ans:

Correlation (Pearson) is also called 'r' or 'Pearson`s R'.
Pearson r measures the strength of the linear relationship between 2 variables. The value of correlation coefficient ranges from -1 to 1.
-1 indicates perfect negative correlation.
1 indicates perfect positive correlation.
-1 or 1 also indicates that all the values of the y variable lie on the line.
The formula to calculate
r = summation ((X - X mean)(Y – Y mean)) / sqrt(summation((X - X mean)^2) * summation((Y – Y mean)^2))

0.8 to 1.0	Very strong relationship
0.6 to 0.8	Strong relationship
0.4 to 0.6	Moderate relationship
0.2 to 0.4	Weak relationship
0 to 0.2	Weak or no relationship

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans:

Scaling means bringing all the continuous variables into smaller scale for easy interpretation of the values. The original ratio of the values doesn't affect.

Scaling is performed for

Ease of interpretation- we can scale all with 0 or 1 to interpret very well.
Faster convergence for gradient descent methods.

Normalized scaling means min-max scaling. It brings the data in the range of 0 to 1.

Minmax Scaling: x = (x-min(x)) / (max(x)-min(x))

Standardisation brings all the data into a standard normal distribution with mean 0 and standard deviation 1.

standardization: x = (x - mean(x)) / SD(x)

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

Ans:

The VIF value is infinite indicates that independent variables have perfect correlation.
VIF (Variance Inflation Factor) is used to identify multicollinearity.
< 5 means no multicollinearity
> 5 should not be ignored
> 10 definitely high correlation.

Measured with VIF = 1 / (1 - R^2).
For perfect correlation R^2 value becomes 1. As R^2 increases VIF value also increases significantly. And with R^2 = 1, the VIF value tends to infinite.

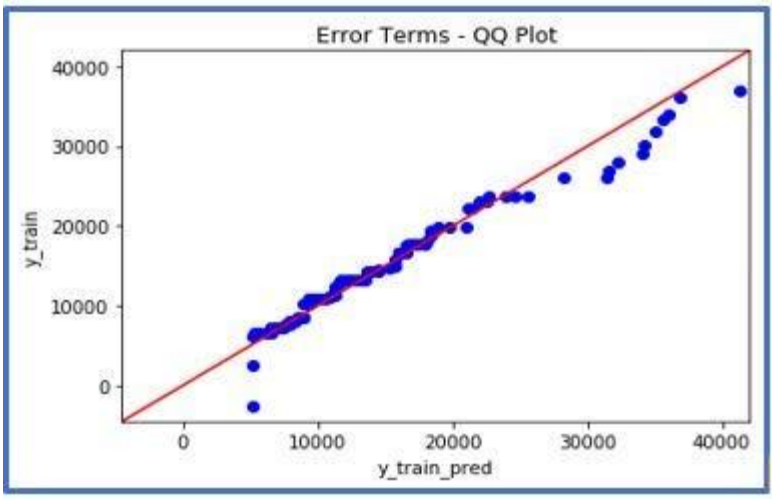
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Ans:

It is also known as quantile-quantile plots. It's a quantiles of sample distribution against quantile of a theoretical distribution. It helps to determine probability distribution like normal, uniform and exponential.
This helps in scenario of linear regression when we have training and test dataset received separately and then we can confirm using Q-Q plot that both the datasets are from populations with same distributions.
Advantage:

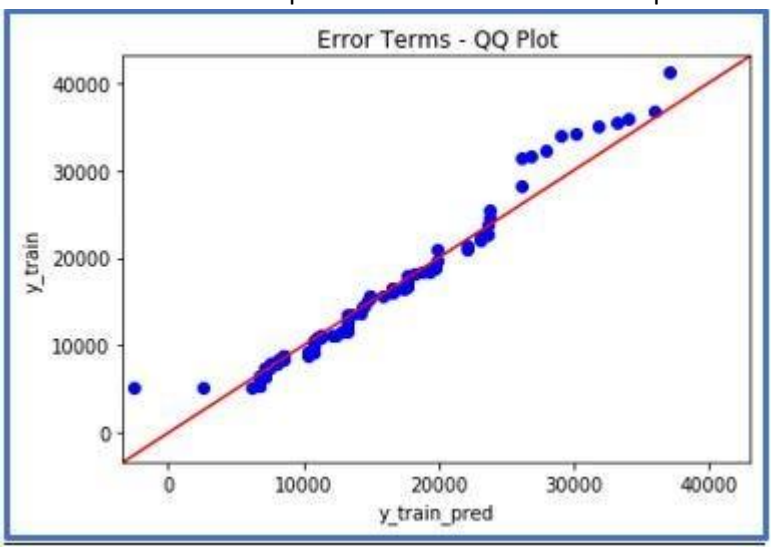
1. It can used with sample sizes also.
2. Many distributional aspects like shifts in location, shifts in scale, changes in symmetry and presence of outliers can also be detected from the plot

Interpretation:

1. Similar distribution: If all points 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 points of quantiles lies away from the straight line at an angle of 45 degree from x-axis.

Python provides qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively.