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)**

- Season : Fall season has higher count of demand when compared to other seasons of summer,winter,spring
- June,July,August,September has higher demand when compared to others months of the year.
- Comparatively, during holidays a bit lower demand is observed.
- There is not much drop or raise during all the days of a week, just a bit lower demand on Sundays. Demand was spread uniformly with minimal change across all days.
- There is no much difference in demand during working day and non-working day.
- Regarding weather situation, cloudy days had more demand than mist and rain. On rainy days it is much lower as expected, since people stay at indoors.

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

During dummy variable creation, it removes the first extra column created thus helps in reducing the correlations created among dummy variables. Default value is false.

e.g. out of k categorical variables, whether to keep k or k-1 variables, drop_first=True will remove first variable.

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

Target variable is count, it is highly correlated with temp and atemp variables which are independent numerical variables.

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

- Error terms being independent
- Equal variance on the predicted and actual data points
- Linear relationship between actual and predcited values.
- Lower and acceptable VIF values (<4)

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

The top 3 features contributing considerably towards the demand of shared bikes are:
Windspeed
Temp
July

General Subjective Questions

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

Linear regression is a Machine Learning Algorithm which is a statistical model to predict the relation between dependant and independent variables. Linear regression makes predictions for continuos (i.e. variables which tend to either increase or decrease) or numeric variables.

E.g. Age, Salary, Temperature, Earnings, Revenue, Sales etc.

The term linear regression itself describes that there would be a linear relationship between a dependant variable say, y and one or more independent variables, say x. Which means how y is increasing or decreasing based on the values of x when they increase or decrease.

Mathematically it is defined as $y = mX + C$

Where y – dependent variable

X – independent variable

M – slope

C – constant

Types of Linear regression –

If one single independent variable is used to analyse and predict the behaviour of the dependent variable it is called simple linear regression.

If more than one single independent variable is used to analyse and predict the behaviour of the dependent variable it is called multiple linear regression.

If the dependent variable increases on Y-axis and independent variable increases on X-axis, then the relationship is called Positive linear relationship.

If the dependent variable decreases on Y-axis and independent variable increases on X-axis, then the relationship is called Negative linear relationship.

Below are the assumptions used while building a Linear Regression model, to ensure we get best result from the dataset.

- Linear relationship between the dependent and independent variables.
- Small or no multicollinearity between the independent variables i.e. multicollinearity means high-correlation between the independent variables which means it is difficult to determine which predictor variable is affecting the target variable.
- Homoscedasticity is when the error term is the same for all the values of independent variables i.e. there is no clear distribution of data.

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

Anscombe's Quartet is a group of four data sets that are nearly identical in statistics involves variance, and mean of all x,y points in all four datasets.

, but have some different patterns in the dataset which means they have very different distributions when plotted on scatter plots.

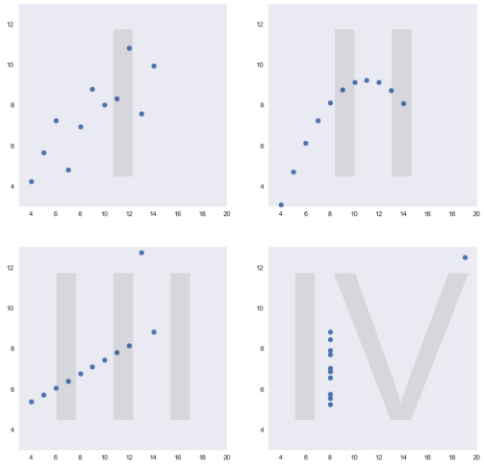
They are used illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. And to see the distribution of the samples that help to identify various anomalies in the data like outliers, spread of the data, linear relationship of the data, etc.

These four below show different patterns:

The column data seems almost similar with same values of mean, variance etc. when calculated. But once they are plotted, they show different patterns,

- 1st plot shows good linear relationship.
- 2nd plot shows no normal distribution.
- 3rd plot shows linear relationship but has outliers.
- 4th plost shows a high value outlier which affects cofficient value.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
0	10	8.04	10	9.14	10	7.46	8	6.58
1	8	6.95	8	8.14	8	6.77	8	5.76
2	13	7.58	13	8.74	13	12.74	8	7.71
3	9	8.81	9	8.77	9	7.11	8	8.84
4	11	8.33	11	9.26	11	7.81	8	8.47
5	14	9.96	14	8.10	14	8.84	8	7.04
6	6	7.24	6	6.13	6	6.08	8	5.25
7	4	4.26	4	3.10	4	5.39	19	12.50
8	12	10.84	12	9.13	12	8.15	8	5.56
9	7	4.82	7	7.26	7	6.42	8	7.91
10	5	5.68	5	4.74	5	5.73	8	6.89



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

The Pearson correlation coefficient is used to know the strength of linear relationship between two data samples of same length.

Pearson's R = covariance(X, Y) / (stdv(X) * stdv(Y))

It is calculated by the covariance of the two variables divided by the product of standard deviation of each data sample.

The above calculation result will help understanding the correlation coefficient relationship.

The coefficient would return a value between -1 & 1 which is the limit of correlation from a negative correlation to a positive correlation.

0 value means no correlation.

A value below -0.5 or above 0.5 indicates a considerable correlation.

Values below those ranges says not so considerable correlation.

E.g. A correlation value of 0.7 means a high level of correlation (i.e. value above 0.5 and below 1.0)

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

Scaling is a data pre-processing step used on independent variables to bring the data to a particular range.
Why scaling means many of the times dataset has values which range from single digits to 6 figures for examples. If scaling is not done the algorithm takes magnitude for analysis and not units, so the model would be terribly be incorrect modelling and is meaningless, so to avoid it we have to bring all the variables to a specific scale with in a range.
Scaling affects only coefficients, not any other parameters like, p-value, t-value, R-squared etc.

Normalization rescales the values into a range of [0,1] or [-1, 1].
Minimum and maximum value of variable values are used for scaling.
It is used when variable values are of different scales.
It is really affected by outliers.

Standardization rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).
Mean and standard deviation is used for scaling.
It is used when we want to ensure zero mean and unit standard deviation.
It is not bounded to a certain range.
It is much less affected by outliers.

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

(3 marks)

If VIF = infinity it is a perfect correlation between two independent variables, corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

To avoid this we need to drop one of the variables from dataset that is causing multicollinearity.
Generally, if VIF > 10 then multicollinearity exists among the variables.

- 1 : not correlated.
- 1 to 5 : moderately correlated.
- 5 : highly correlated.

To handle multicollinearity remove some of the highly correlated independent variables or combine linearly independent variables.

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

(3 marks)

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other to find if two sets of data come from the same distribution.
A quantile is a fraction where certain values fall below that quantile.
E.g. Median is a quantile where 50% of the data fall below that point and 50% lie above it.

It is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.

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

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios: If two data sets —

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes and tail behavior