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)

1. Average demand is highest in 'fall' followed by 'summer' then 'winter'. Demand is lowest in Spring and the drop looks more than 50%.

2. Average demand in 2019 has surged by around 50% compared to 2018.

3. There is gradual increase in average demand from January to September and then drops gradually may be due to holiday season and/or snow/rain fall.

4. There is low average demand on holiday than not-holiday, the drop looks around 30%.

5. Average demand across week is stable.

6. Average demand is same on working and non-working days though some non-working days there is low demand(lower quartile is lower than \`working\` in box-plot comparison).

7. The average demand is high when the weather is clear and the drop in demand looks more than 50% when there is light rain/snow.

8. The average demand on non-workingday is almost same as on workingday if it is not-holiday. As per the dictionary a non-working day which is not a holiday is a weekend so the demand is almost same on weekends as is on workingday.

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

When creating dummy variables for a categorical variable with k categories, k different columns are created each one with binary values. If we analyse all these k columns it can be seen that first column can be explained or predicted by remaining k-1 columns which means that first variable is redundant and will show multicollinearity if included in modelling. To avoid multicollinearity between dummy variables and a redundant column in dataset we use drop_first=True to drop the column.

Consider example of categorical variable `furnishingstatus` which has three levels `furnished`, `unfurnished` and `semi-furnished` when we create dummy variables 3 different variables will be created.

furnished	unfurnished	semi-furnished
1	0	0
1	0	0
0	1	0
0	0	1

We can see that the type of `furnished` can be identified with other two columns where
`00` will correspond to `furnished`
`10` will correspond to `unfurnished`
`01` will correspond to `semi-furnished`

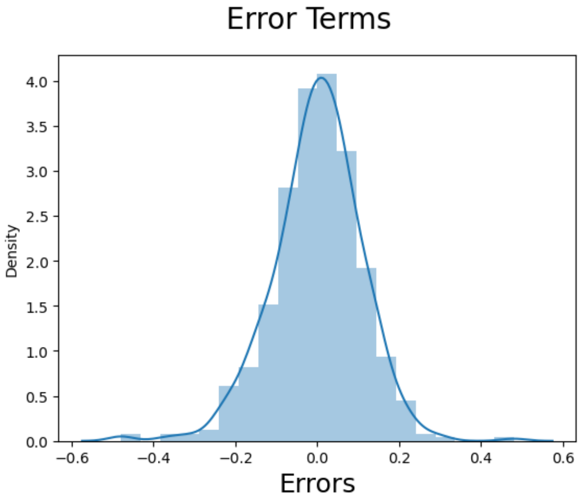
If we drop ‘furnished’ column the other 2 will still have the info of the dropped column hence ‘furnished’ is redundant.

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

’temp’ and ‘atemp’ both have the highest correlation among other numerical variables. The correlation coefficient with ‘cnt’ is observed as 0.63 which is significant in the heat-map.

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

   The assumption that “Error terms are normally distributed with mean zero” was validate by observing the histogram of errors. Errors are the difference between predicted values and actual values of training data. It was observed that the residuals are normally distributed with a mean 0 as seen in the below figure.



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

   Based on the final model the top 3 features are;

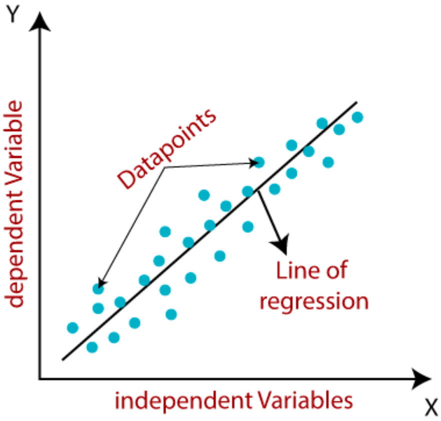
   1. ‘Light Rain’ (Light rain/snow weather) which has negative correlation with demand for bikes.
   2. ‘spring’ (Spring season) which has negative correlation with demand for bikes.
   3. ‘yr’ (Year) which has positive correlation with demand for bikes.

General Subjective Questions

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

In Machine Learning, Linear Regression algorithm is used for predicting a output variable based on one or more input variables. The output variable is also called as dependent variable or target variable and the input variables are also called as independent variables or predictor variables.

Linear regression is used for predicting continuous/numeric variable. It requires that the target variable has a linear relationship with the predictor variables and that the predictor variables are independent.



Linear regression models can be classified into two types depending upon the number of independent variables as;

Simple linear regression: When the number of independent variables is 1.

The general equation is: y = mx + b

y is the dependent variable (target)
x is the independent variable (feature)
m is the slope of the line
b is the y-intercept.

Multiple linear regression: When the number of independent variables is more than 1.

The general equation is: y = b0 + b1x1 + b2x2 + …….bnxn

y is the dependent variable (target)
b0 is the y-intercept, the value of y when all x values are 0.
b1, b2, b3…..bn are the coefficients associated with the respective independent variables x1, x2, x3…..xn.
x1, x2, x3….xn are the independent variables (features).

Linear regression is a Supervised Learning method and the goal is to find the best fit linear model by learning from the training data.

The objective of the linear regression algorithm is to find the values of m, b (or b1, b2, b3…..bn) that minimise the difference between the predicted values and the actual values in the training data.

It can be found by minimising the difference/error between the predicted values and the actual values in the training data. This is typically achieved by using the method of Ordinary Least Squares(OLS), which involves minimising the sum of the squared differences between the predicted and actual values. Optimisation techniques like differentiation and Gradient Descent are used.

The strength of a linear regression model or how well the line is fit is mainly determined by a metric R^2 which is defined as $R^2 = 1 - (RSS / TSS)$

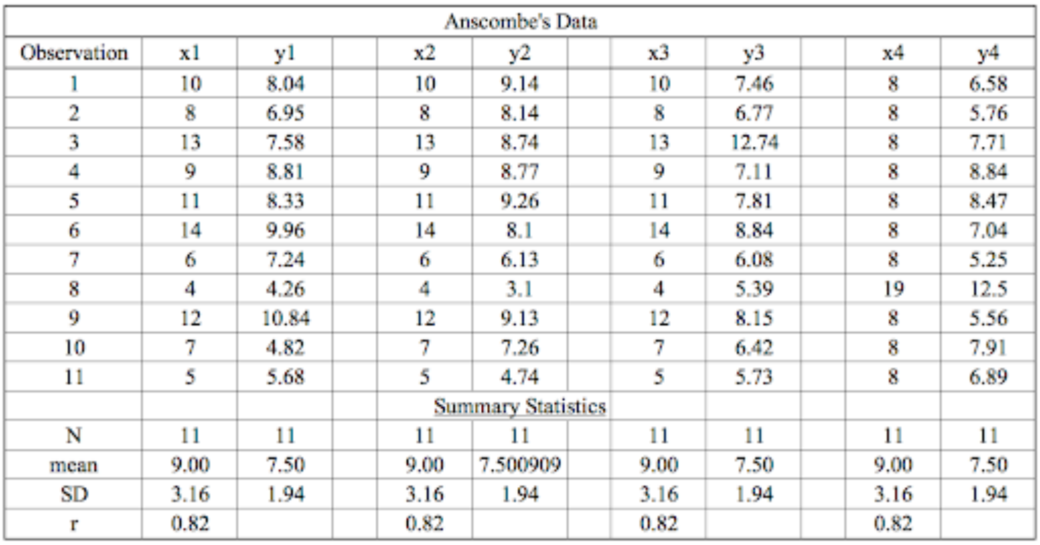
RSS: Residual Sum of Squares
TSS: Total Sum of Squares

R^2 varies between values 0 and 1 and value nearer to 1 is considered better for example a 0.75 value means the model explain 75% of variance in target variable which is considered a good fit.

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

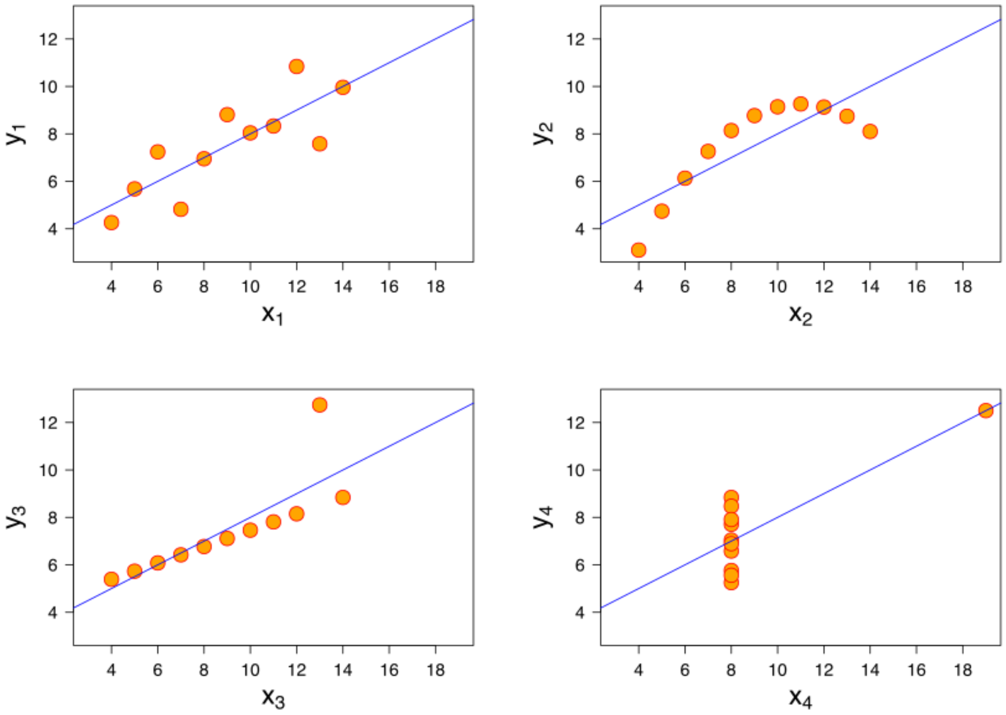
Anscombe's quartet is a group of datasets (x, y) that have nearly identical simple descriptive statistics like mean, standard deviation, and regression line, but have very different distributions and appear very different when graphed.

They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analysing it, and the effect of outliers and other influential observations on statistical properties.



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	



As seen from above figures the statistical information for the four data sets are approximately similar but each are qualitatively different when plotted and distribution does not fit a common model.

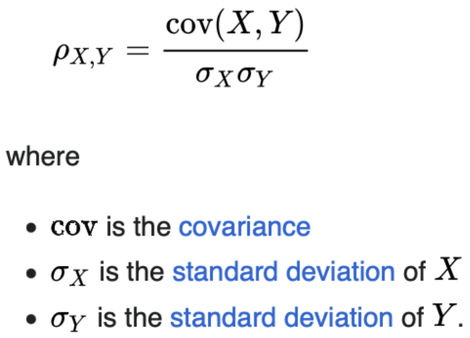
Anscombe's quartet is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties.

3. What is Pearson’s R? (3 marks)

Parson’s R is a correlation coefficient that measures linear correlation between two continuous variables and is commonly used in Linear regression. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables that are measured on the same interval.

Pearson coefficients range from +1 to -1, with +1 representing a positive correlation(when one increases other increases), -1 representing a negative correlation(when one increases other decreases), and 0 representing no linear relationship.

It is defined as the covariance of the two variables divided by the product of their standard deviations.



$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y.

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

In Machine Learning modelling, feature scaling is a data preprocessing technique used to transform the values of features or variables in a dataset to a similar scale. In general scaling refers to putting the feature values into the same range.

Scaling is performed in datasets containing features that have different ranges, units of measurement, or orders of magnitude to ensure that no single feature dominates the distance calculations in an algorithm and improve performance. Scaling also avoids biased models and optimisation algorithms converge faster.

Normalised Scaling also known as Min-Max Scaling scales the features to a specific range, usually between 0 and 1.

It is useful when the features have different ranges, and we want to ensure that they are all on a similar scale. It is typically used when the distribution of data is approximately uniform.

The normalised value is calculated by following formula;

$$x_{\mathrm{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Standardised Scaling transforms the features to have a mean of 0 and a standard deviation of 1. It is also known as Z-score normalisation.

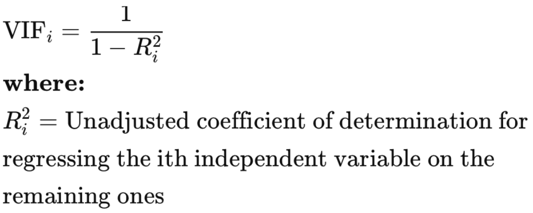
It is useful when the features have different units and when the distribution of your data may not be normal. It's also commonly used when applying machine learning algorithms that rely on gradient-based optimisation. It can make optimisation algorithms converge faster.

The standardised value is calculated by following formula;

$$x_{\mathrm{stand}} = \frac{x - \mathrm{mean}(x)}{\mathrm{standard\ deviation}\ (x)}$$

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

A variance inflation factor (VIF) is a measure of multicollinearity among the independent variables in a multiple regression model. The greater the VIF, the higher the degree of multicollinearity.



$$\text{VIF}_i = \frac{1}{1 - R_i^2}$$

where:

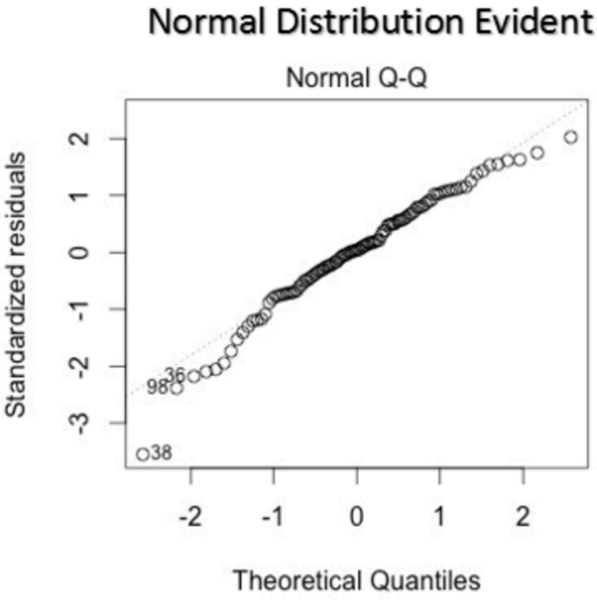
R_i^2 = Unadjusted coefficient of determination for regressing the ith independent variable on the remaining ones

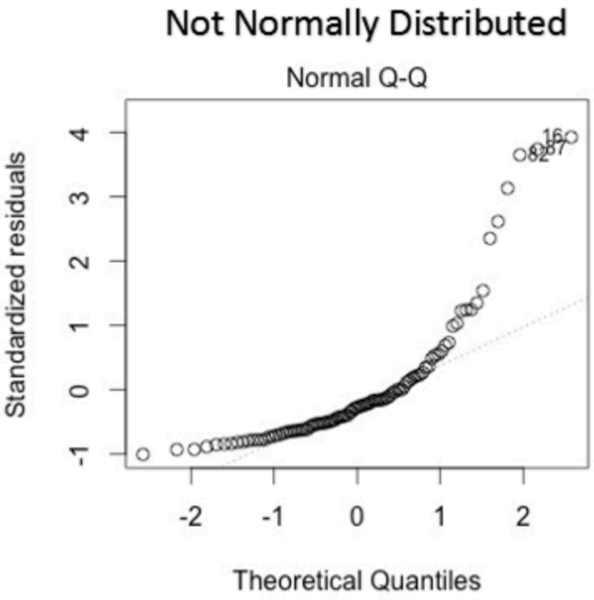
VIF is calculated using above formula so when R² is 1 in the above formula then VIF will be infinity. R² is the measure of the strength of collinearity with other predictor variables. When R² value is 1 i.e. a perfect collinearity means the variance in the variable is explained 100% by the the other variable so in such case of perfect collinearity value of VIF will become infinite.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

   The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. It is a probability plot for comparing two probability distributions by plotting their quantiles against each other.

   In Linear regression a Q-Q plot can be used to check if the residuals of the model are normally distributed, which is an assumption for many parametric tests and confidence intervals. It can also be used to check if the residuals have a constant variance, which is an assumption for the homoscedasticity of the model. To do this, you need to create a Q-Q plot for the residuals of the model and compare them with the normal distribution.


Normal Distribution Evident
Normal Q-Q
Standardized residuals
Theoretical Quantiles


Not Normally Distributed
Normal Q-Q
Standardized residuals
Theoretical Quantiles