Assignment-based Subjective Questions

QUESTION 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 1:

a. *We see that during fall followed by summer has high demand in shared bikes as compared to winter and spring*
b. *Also, we see high demand in 2019 as compared to 2018 year*
c. *We saw during the month from May till October the demand is high which keeps on increasing till October and started to decrease from November*
d. *We see little less demand when it is a holiday as compare to regular days*
e. *We didn't see much pattern during the days of the week the count seems to be pretty much similar, same with working days*
f. *We do see that weather condition plays important role in the demand as when the weather is good or cloudy the demand is more and during bad days (rainy, snowy) the demand decreases.*

QUESTION 2.

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

Answer 2:

i. ***drop_first=True*** *removes the first column which is created for the first unique value of a column. It is important, as it helps in reducing the extra column created during dummy variable creation*

QUESTION 3.

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

Answer 3:

i. *Looking at the pair plot we see the highest Corelation of* ***CNT(target variable)*** *is with temperature (****TEMP****) variable.*

QUESTION 4.

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

Answer 4:

a. *We plotted Heatmap graph in order to check if there is any multi collinearity in the variables selected after building the model and found that there were no extreme value that could prove multi collinearity existence.*
b. *We also check the VIF and p value and made sure no high values of p exists so that we have only significant variables and also no value higher than 4 in VIF to avoid multi collinearity.*
c. *We plot the actual and predicted graph in order to check if the predicted value is able to follow the actual value pattern.*
d. *At the end we did the Residual Analysis so that we could check if the residual left is normally distributed around the mean and there are no hidden patterns in the error.*

QUESTION 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 5:

a. *Temperature (highly corelated – 0.3577)*
b. *Year (higher demand in 2019 as compare to 2018)*
c. *Season (Fall has high demand and Spring has least)*

General Subjective Questions

QUESTION 1.

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

Answer 1:

***Linear Regression** is a machine learning algorithm based on supervised learning.*

a. *It performs a regression task. Regression models a target prediction value based on independent variables.*
b. *It is mostly used for finding out the relationship between variables and forecasting.*
c. *Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.*
d. *Aims is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized.*

Linear Regression is of two types: Simple and Multiple.

a. *Simple Linear Regression is where only one independent variable is present and the model has to find the linear relationship of it with the dependent variable.*
b. *Multiple Linear Regression there are more than one independent variables for the model to find the relationship.*

*The **basic assumptions** of Linear Regression are as follows:*

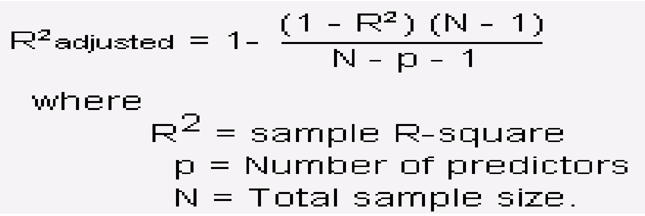
i. *Linearity: It states that the dependent variable Y should be linearly related to independent variables. This assumption can be checked by plotting a scatter plot between both variables.*
ii. *Normality: The X and Y variables should be normally distributed. Histograms, KDE plots, Q-Q plots can be used to check the Normality assumption.*
iii. *Homoscedasticity: The variance of the error terms should be constant i.e., the spread of residuals should be constant for all values of X. This assumption can be checked by plotting a residual plot. If the assumption is violated then the points will form a funnel shape otherwise, they will be constant.*
iv. *Independence/No Multicollinearity: The variables should be independent of each other i.e.; no correlation should be there between the independent variables. To check the assumption, we can use a correlation matrix or VIF score. If the VIF score is greater than 5 then the variables are highly correlated.*
v. *Error terms should be normally distributed. Q-Q plots and Histograms can be used to check the distribution of error terms.*
vi. *No Autocorrelation: The error terms should be independent of each other. Autocorrelation can be tested using the Durbin Watson test. The null hypothesis assumes that there is no autocorrelation. The value of the test lies between 0 to 4. If the value of the test is 2 then there is no autocorrelation.*

Evaluation Metrics for Regression Analysis

1. ***R squared**: The value of R squared lies between 0 to 1, the value closer to 1 the better the model.*

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

2. ***Adjusted R squared:** It only considers the features which are important for the model and shows the real improvement of the model. It is always lower than R2.*



$$R^2_{\text{adjusted}} = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

where

R^2 = sample R-square

p = Number of predictors

N = Total sample size.

3. ***Mean Squared Error (MSE):** Another Common metric for evaluation is Mean squared error which is the mean of the squared difference of actual vs predicted values.*

$$MSE = \frac{1}{n}\Sigma\left(\underbrace{y - \widehat{y}}_{\text{The square of the difference between actual and predicted}}\right)^2$$

4. ***Root Mean Squared Error (RMSE):*** *It is the root of MSE i.e Root of the mean difference of Actual and Predicted values. RMSE penalizes the large errors whereas MSE doesn't**.***

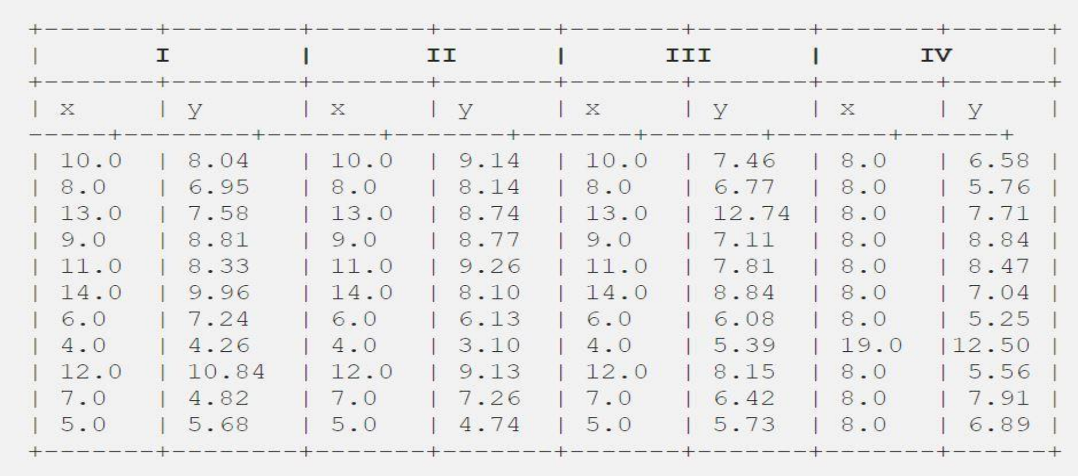
$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

QUESTION 2.

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

Answer 2:

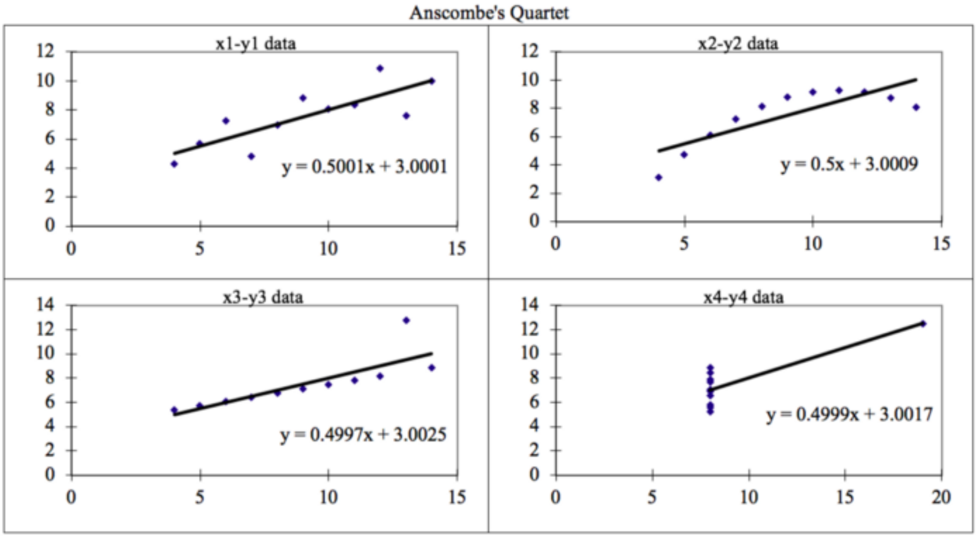
<u>Anscombe's Quartet</u> *can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter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



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

*The **four datasets** can be described as:*

i. *<u>Dataset 1</u>: this fits the linear regression model pretty well.*
ii. *<u>Dataset 2</u>: this could not fit linear regression model on the data quite well as the data is non-linear.*
iii. *<u>Dataset 3:</u> shows the outliers involved in the dataset which cannot be handled by linear regression model*
iv. *<u>Dataset 4:</u> shows the outliers involved in the dataset which cannot be handled by linear regression model*



Application:
The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

QUESTION 3.

What is Pearson's R? (3 marks)

Answer 3:

Correlation *is a bi-variate analysis that measures the strength of association between two variables and the direction of the relationship.*

i. *In terms of the strength of relationship, the value of the correlation coefficient varies between +1 and -1. A value of ± 1 indicates a perfect degree of association between the two variables.*
ii. *As the correlation coefficient value goes towards 0, the relationship between the two variables will be weaker.*
iii. *The direction of the relationship is indicated by the sign of the coefficient; a + sign indicates a positive relationship and a - sign indicates a negative relationship.*

Four types of correlations*:*

i. *Pearson correlation*
ii. *Kendall rank correlation*
iii. *Spearman correlation*
iv. *Point-Biserial correlation.*

Pearson R Correlation

Pearson correlation coefficient is a measure of the strength of a linear association between two variables — denoted by r.

i. *For the Pearson r correlation, both variables should be normally distributed. i.e the normal distribution describes how the values of a variable are distributed. This is sometimes called the 'Bell Curve' or the 'Gaussian Curve'*
ii. *There should be no significant outliers.*
iii. *Each variable should be continuous*
iv. *The two variables have a linear relationship. Scatter plots will help you tell whether the variables have a linear relationship*
v. *The observations are paired observations. That is, for every observation of the independent variable, there must be a corresponding observation of the dependent variable*
vi. *The error term is the same across all values of the independent variables. If the points lie equally on both sides of the line of best fit, then the data is homoscedastic.*

QUESTION 4.

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

Answer 4:

Scaling:

This means that you're transforming your data so that it fits within a specific scale, like 0-100 or 0-1.

i. *You want to scale data when you're using methods based on measures of how far apart data points, like support vector machines, or SVM or k-nearest neighbours, or KNN.*
ii. *With these algorithms, a change of "1" in any numeric feature is given the same importance.*
iii. *By scaling your variables, you can help compare different variables on equal footing.*

For example, *you might be looking at the prices of some products in both Yen and US Dollars. One US Dollar is worth about 100 Yen, but if you don't scale your prices methods like SVM or KNN will consider a difference in price of 1 Yen as important as a difference of 1 US Dollar! This clearly doesn't fit with our intuitions of the world. With currency, you can convert between currencies. But what about if you're looking at something like height and weight? It's not entirely clear how many pounds should equal one inch (or how many kilograms should equal one meter).*

Why Scaling is performed:

Machine learning algorithms like linear regression, logistic regression, neural network, etc. that use gradient descent as an optimization technique require data to be scaled. Take a look at the formula for gradient descent below:

Gradient descent formula

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

i. *The presence of feature value X in the formula will affect the step size of the gradient descent. The difference in ranges of features will cause different step sizes for each feature.*

ii. *To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model.*

iii. *Having features on a similar scale can help the gradient descent converge more quickly towards the minima*

Normalization *is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.*
Here's the formula for normalization:

Normalization equation

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

i. *When the value of X is the minimum value in the column, the numerator will be 0, and hence X' is 0*

ii. *On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X' is 1*

iii. *If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1*

Standardization *is another scaling technique where the values are centred around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.*

Here's the formula for standardization:

Standardization equation

$$X' = \frac{X - \mu}{\sigma}$$

i. *Feature scaling: Mu is the mean of the feature values*

ii. *Feature scaling: Sigma is the standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range.*

QUESTION 5.

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

Answer 5:

This shows a perfect correlation between two independent variables.

i. *In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.*
ii. *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 (which show an infinite VIF as well)

QUESTION 6.

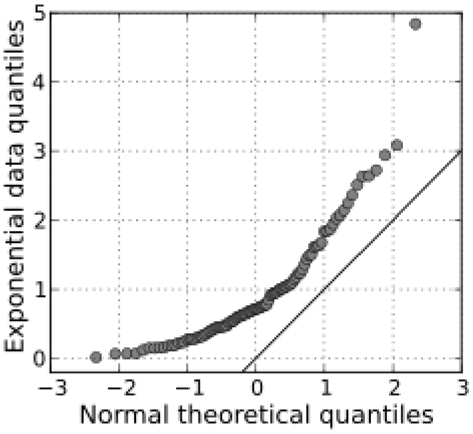
What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer 6:

Q-Q Plots (Quantile-Quantile plots) *are plots of two quantiles against each other.*

i. *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.*
ii. *The purpose of Q Q plots is to find out if two sets of data come from the same distribution.*
iii. *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.*

A Q Q plot showing the 45-degree reference line:



1. *If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x.*
2. *If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x.*
3. *Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.*

Importance

A Q–Q plot 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.