**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)**

A a box-plot with a subplot was used to see the effect of the categorical variables on ‘cnt’, ‘registered’ and ‘casual’. The following details were observed

#### ‘weathersit’ seems to have a negative co-relation with ‘cnt’.

‘workingday’ does not seem to impact the count.

‘holiday’ has a negative co-relation. Slightly smaller numbers.

‘mnth’ - There is significantly less count over the initial and final months of the year

‘yr’ - there has been significant increase in the second year

‘season’ has lot of correlation - significantly less in Spring and winter. Fall and Summer seeing higher count

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

drop\_first=True makes sure that a categorical variable with n levels is represented by n-1 dummy variables and not n variables. N variables will have redundant information.

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

‘cnt’ has highest co-relation with ‘temp’ followed by ‘atemp’.

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

Following are the assumptions in Linear Regression

1. Linear relationship between X and Y – A plot of the predicted values in the training set shows the linear relationship between the predictor variables and the target variable.

2. Error terms are normally distributed (not X, Y)

A plot of the residuals showed that the error terms were normally distributed around a mean of 0

3. Error terms are independent of each other

A plot of all the error terms shows that error terms are distributed around 0.

4. Error terms have constant variance (homoscedasticity)

From the scatter plot of the predicted value for y, it is visible that the predicted value does not have too much variance.

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

Top three features are

atemp

yr

weathersit

This was concluded based on the value of the co-efficients derived from the model.

**General Subjective Questions**   
1. **Explain the linear regression algorithm in detail. (4 marks)**

Linear regression explains the relationship between a dependent and independent variable using a straight line.The standard equation of a linear regression line is given as y = β0 + β1\*X

If the effect of one predictor variable on one target variable is calculated then it is called Simple Linear Regression. β0 is the intercept and β1is the slope of the line. Here if X is increased by 1 unit then y increases by β1.

If the effect of more than one predictor variable on the target variable is calculated then it is called Multiple Linear Regression. Here β1 is the change in Y when X is increased by one unit, assuming all the other values remain the same.

Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable:

Fig 4 .The idea behind building a linear regression model is to determine the values of β0 and β1 such that the residual sum of squares is minimized.

Residual error is given as

e = Ypred – Y

RSS(Residual Sum of Squares) = e12 + e22……..en2

RSS = (Y1- β0 - β1\*X1)2+……….(Yn- β0 – βn\*Xn)2

The best fit for a line is got by minimising the value of RSS.

RSS is an absolute value, hence if there is a change in the scale of the predictor variables then the value of RSS also changes. Total Sum of Squares which is a relative value will be a better indication.TSS is the total sum of the errors of the data points from mean of the response variable. RSS/TSS normalizes a model and indicates how well the variance is explained.

Two metrics used the define the strength of a Linear Regresion Model is

R2 It explains how much variance in the value of y is explained by the model

It takes a value between 0 and 1 where 1 is the best fit and 0 is the worst. Higher value of R2 means a better model.

RSE(Residual Standard Error)

Assumptions made in inferences from Linear Regression Models are as follows

X and y have a linear relationship

The error terms or residuals are normally distributed

Errors are independent of each other

Errors have a constant variance

In inferences from Multi Linear Regressions have the following considerations

Models may become too complex and overfit due to the presence of multiple features

Some of the features may be associated. This will make the coefficient values are affected by this and the inferences might not be correct.

Feature selection also is an important aspect

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

Anscombe’s quartet explains the importance of data visualization. It comprises of four data sets that have similar statistics. They seem to a have the same mean and variance. However when plotted on a graph they show different characteristics

Chart, scatter chart

Description automatically generated

In the graph shown above we can see that all the 4 data sets are

represented by the same line when we plot the individual data points, some new characteristics are seen. You can see that only the third graph captures the variance in the data effectively.

3. **What is Pearson’s R? (3 marks)**

Correlation coefficients are used to measure the relationship between two variables. It shows how strongly or weekly related the variables are. There are several types of correlation coefficients. Pearson’s R is a popular one used in Linear Regression.

Correlation coefficients usually return a value between -1 and 1.

1 indicating a strong positive relationship

-1 indicating a strong negative relationship

0 indicating no relationship

Pearson’s R is used to measure the relationship between continuous variables.   
**Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations**.

For example, a child's height increases with his increasing age (different factors affect this biological change). So, we can calculate the relationship between these two variables by obtaining the value of Pearson's Correlation Coefficient r.

There are certain requirements for Pearson's Correlation Coefficient:

* Scale of measurement should be interval or ratio
* Variables should be approximately normally distributed
* The association should be linear
* There should be no outliers in the data

The formula is given by

Logo

Description automatically generated with medium confidence

Where

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores

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

When different features in a model have different scales, the coefficients also vary a lot and don’t give a good idea of how they effect the result.

This will help in

1. Ease of interpretation
2. Faster convergence for gradient descent models

There are two popular methods of scaling

1. Standardizing - The variables are scaled in such a way that their mean is zero and standard deviation is one.

x = x-mean(x)/sd(x)

1. Max Min Scaling - The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

x = x-min(x)/max(x)-min(x)

Scaling does not affect the statistic quality of a model. Only the coefficients are changed.

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

If there is perfect correlation between two variables then VIF will be infinite. The variable can be perfectly represented by a combination of other variables.

When a VIF is 5. From the formula VIF = 1/1-R2,

We get R2 = 0.8

80% of the variance is explained by the other variable. Hence a very high value like infinity shows that they are almost the same representation.

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

Q-Q plot or Quantile-Quantile plots are used to find the type of distribution for a random variable. If we assume that the dependent variable is Normally distributed then we can use a Q-Q Plot to check that assumption. It is a visual check.

If observations are normally distributed then Q-Q Plot will result in a straight line. Many statistical inference procedures assume that we are sampling from a normally distributed population. Hence assessing normality is important.

Using a QQ Plot , you can assess if two sets of data came from the same distribution. A 45o line is plotted on the QQ plot. If the two sets of data come from the same distribution, the points will fall on the reference line.  
(3 marks)