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)

I have done this Categorical Bivariate analysis with dependent variable(cnt) using boxplot. Below are my analysis report for the same

1. Looks like fall is having high demand of bike sharing and spring is having the lowest
2. In year 2 the demand of the bike sharing increased significantly
3. Demand increased continuously from Jan to Jun and it was highest in Sept then it gradually decreasing till Dec
4. First 10 days of the month demand is little high then slight low in 2nd and then 3rd 10 days of the month
5. Holidays demand is less may be people not going to office and want spend time at home
6. Demand is little high during Wednesday, Thursday than other working days. Its may be due to hybrid way of working .People may work from office more on Wednesday and Thursday then prefer other days to work from home. Also Saturday is having more demand than Sunday as they want to take rest and be with family.
7. On working days demand is slightly high
8. Good and clear weather is having high demand of bike sharing than other

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

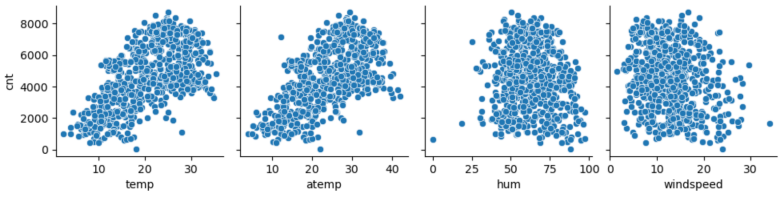
When we are dealing with categorical variables during Model creation we need to create dummy variable and to avoid intercorrelation we need to remove one dummy variable column. If we have k levels in a categorical variable then we need to create k-1 dummy variable column. So, to handle this at run time instead of doing it manually later we are using drop_first=True.

e.g. `wsit_dum = pd.get_dummies(ds["weathersit"], drop_first=True, dtype=int)`

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

Both temp and atemp are looking highly corelated with cnt from the pair-plot graph. But from the corelation matrix atemp is highest corelation with cnt.

```
In [16]: sns.pairplot(x_vars=ds[cont_cols], y_vars=ds[target], data=ds)
Out[16]: <seaborn.axisgrid.PairGrid at 0x12d0149d0>
```




```
In [18]: corr_ = ['temp', 'atemp', 'hum', 'windspeed', 'cnt']
         ds[corr_].corr()
```

Out[18]:

	temp	atemp	hum	windspeed	cnt
temp	1.000000	0.991696	0.128565	-0.158186	0.627044
atemp	0.991696	1.000000	0.141512	-0.183876	0.630685
hum	0.128565	0.141512	1.000000	-0.248506	-0.098543
windspeed	-0.158186	-0.183876	-0.248506	1.000000	-0.235132
cnt	0.627044	0.630685	-0.098543	-0.235132	1.000000

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

Linear Regression model is validated with below assumptions

1. Error terms are normally distributed around zero
2. Multicorelation check
3. Linear relation with dependent variable

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

Below features are contributing significantly high towards explaining the demand

1. temp
2. winter
3. sep

General Subjective Questions

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

Linear regression is a statistical method used for modelling the relationship between a dependent variable and one or more independent variables. LR Model helps us to find the best-fitting linear relationship that describes the data. It tries to establish a linear equation that predicts the dependent variable based on the values of the independent variables.

Linear regression equation for a single independent variable is: $y = Mx+b$

- y is the dependent variable
- x is the independent variable
- M is the slope of the line
- b is constant

It's called Simple Linear Regression(SLR)

In the case of multiple independent variables, the equation becomes:

$y=b+M1x1+M2x2+…+Mnxn$

- y is the dependent variable
- $x1, x2,…xn$ are the independent variable
- $M1, M2,…Mn$ are the slope for $x1, x2,…xn$ *respectively*
- b is constant

It's called Multi Linear Regression(MLR)

These linear relationship could be of positive or negative corelation

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

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

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

Pearson's correlation coefficient, often denoted as r, is a measure of the strength and direction of a linear relationship between two variables. It ranges from -1 to 1, where:

- $r=1$: Perfect positive linear correlation
- $r=-1$: Perfect negative linear correlation

- *r*=0: No linear correlation

The formula for Pearson's correlation coefficient between two variables *X* and *Y* is given by:

r = ∑(Xi – Xmean)(Yi – Ymean)/sqrt(∑(Xi – Xmean)^2 ∑(Yi – Ymean)^2)

Here,
Xi and *Yi* are the individual data points
Xmean and Ymean are the means of *X* and *Y* respectively, and the summation is done over all data points.

Pearson's correlation coefficient measures the strength of a linear relationship. If *r* is positive, it indicates a positive linear correlation (as one variable increases, the other tends to increase). If *r* is negative, it indicates a negative linear correlation (as one variable increases, the other tends to decrease). If *r* is close to 0, it suggests a weak or no linear correlation.

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

Scaling is an important technique which is used in Modelling process before training the data. With this technique we can bring down the high values of features to a comparable scale. E.g. Let one feature values is ranging in millions and another in 0 to. 2 then this is not a comparable scale. To make is comparable we need take help of scaling.

There are n number of techniques are available but we are mainly using below

1. Normalization

Also known as min-max scaling or min-max normalization, it is the simplest method and consists of rescaling the range of features to scale the range in [0, 1]. The general formula for normalization is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

2. Standardization

Feature standardization makes the values of each feature in the data have zero mean and unit variance. The general method of calculation is to determine the distribution mean and standard deviation for each feature and calculate the new data point by the following formula:

$$x' = \frac{x - \bar{x}}{\sigma}$$

We can either manipulate manually with the code or can use sklearn library for this.

1. sklearn.preprocessing.**StandardScaler**
2. sklearn.preprocessing.**MinMaxScaler**

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

VIF is used to find corelation between the independent variables. The measure of VIF is as below

VIF = R / 1-R

If two independent variables are perfectly corelated then then their R values becomes 1 as a result VIF value becomes infinity. For our case on the first model budling we got below variables have VIF as infinity **holiday, workingday, sat and sun**

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 plot is a graphical method for determining whether two samples of data came from the same population or not. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value.

Usage:
The Quantile-Quantile plot is used for the following purpose:

- Determine whether two samples are from the same population.
- Whether two samples have the same tail
- Whether two samples have the same distribution shape.
- Whether two samples have common location behaviour.