Assignment-based Subjective Questions

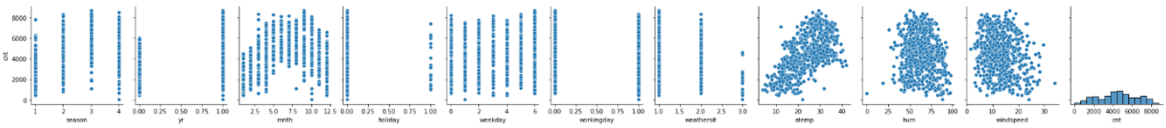
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
(3marks)

Answer : From the final expression we deduce,
cnt = 0.1992 + 0.2363 * yr - 0.0606 * holiday + 0.0546 weekday + 0.0194 workingday + 0.4657 atemp - 0.1381 windspeed - 0.0795 cloudy - 0.2802 scattered - 0.1013 spring + 0.0251 summer + 0.0603 winter,
we understand that the categorical variables such as workingday, the season, as well as the year seems to have significant impact on the dependent variables. When variables like workingday, year seems to have huge positive impact on the demand, season scattered clouds, spring etc seem to negatively affect the demand.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
Answer : drop_first=True helps in reducing the correlation or collinearity. This is possible because if we have 3 variables to be expressed in terms of 1 and 0, its possible to express it using 2 variables as if both are 0, it represents the third variable. Hence, this command helps to reduce the extra variable.

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

Answer : temp/atemp seems to have the highest correlation with the target variable



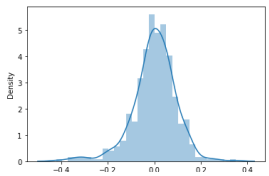
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer : Validation of the assumptions of Linear Regression after building the model was performed using the below steps :

a. To validate if the residual is evenly spread across 0, we used the distplot of the residual (where residual = y_train-y_train_pred)

```
res = y_train-y_train_pred
sns.distplot(res)
```

```
/Users/rahul.kunjumon/MSc/PGIIITB/anaconda3/lib/pyt
a future version. Please adapt your code to use eit
  warnings.warn(msg, FutureWarning)
<AxesSubplot:ylabel='Density'>
```



b. To validate if there is a linear relationship between the dependent and the independent variable, we used the scatter plot. We drew the scatter plot to identify and confirm on the first assumption.

c. We assume that there is a little or no multicollinearity in the data. To validate this, we use the VIF factor and eliminate all the variables with VIF greater than 5.

]:

	Features	VIF
4	atemp	5.17
5	windspeed	4.53
3	workingday	3.15
2	weekday	3.08
8	spring	2.23
0	yr	2.06
9	summer	1.85
10	winter	1.73
6	cloudy	1.54
1	holiday	1.09
7	scattered	1.08

]:
```
### max p value assumed was 0.05 and VIF was 5. a
```

d. We assume that there is little or no autocorrelation in the data. For this we validate the Durbin-Watson test. Durbin-Watson's d tests the null hypothesis that the residuals are not linearly auto-correlated. We assume as a rule of thumb values of $1.5 < d < 2.5$ show that there is

no auto-correlation in the data. Obtained value was 2.041. For Homoscedasticity, we again used a scatter plot

Omnibus:	73.728	**Durbin-Watson:**	2.041
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	187.880
Skew:	-0.732	**Prob(JB):**	1.59e-41
Kurtosis:	5.588	**Cond. No.**	19.6

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 : Top 3 features contributing significantly to the demand of shared bikes were :

a. Atemp. With a highest positive slope of 0.4657
b. Scattered Rain/Snowfall with negative slope of 0.2802
c. Year, with a positive slope of 0.2363

General Subjective Questions

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

Answer : Linear regression is an analysis that evaluates whether one or more predictor variables explain the dependent variable using a straight line.

Two types of linear regression :

Simple linear regression : explains the relationship between a dependent variable and one independent variable using a straight line expressed by $Y = \beta_o + \beta_1 X$

Multiple linear regression : the relationship between one dependent variable and several independent variables expressed by $Y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta pXp + c$

We assess the strength of the linear regression model can be assessed using 2 metrics:

1. R^2 or Coefficient of Determination
2. Residual Standard Error (RSE)

While performing linear regression, we keep some Assumptions, namely :

were:

1. Linear relationship between X and Y
2. Error terms are normally distributed (not X, Y)
3. Error terms are independent of each other
4. Error terms have constant variance (homoscedasticity)

We finally use these parameters to assess a model are:

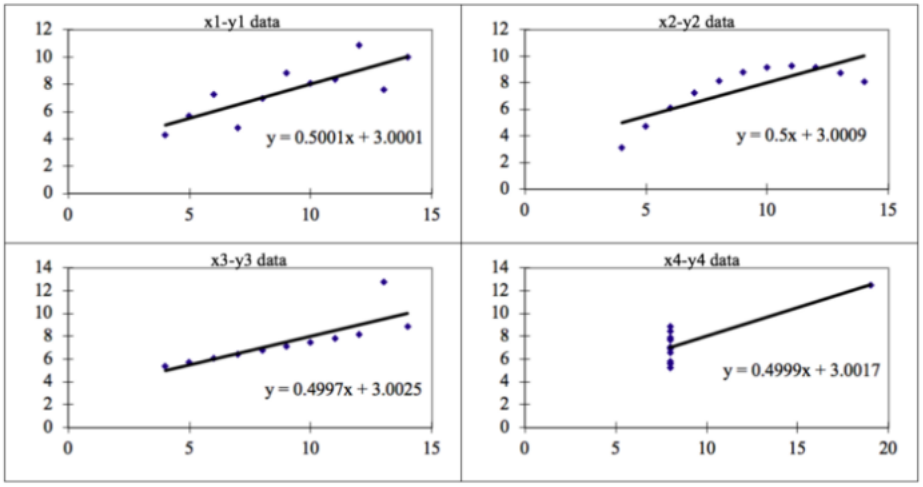
1. t statistic: Used to determine the p-value
2. F statistic: Used to assess whether the overall model fit is significant or not
3. R-squared: the R-squared value tells how well the straight line describes the variance in the data

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

Answer : Anscombe's Quartet can be defined as a group of four data sets which are nearly identical, but there are some speciality in the dataset that decieves the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. Francis Anscombe illustrated

the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets. But, wouldn't help us concluding with the model. The four datasets can be described as: Dataset 1: this fits the linear regression model pretty well.

Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.



** Image and data as perceived and understood online.

Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model

Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

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

Answer : Pearson's correlation coefficient is the statistics that evaluates the statistical relationship between two continuous variables.
It is known as the best method of measuring the association between variables because it is based on the method of covariance. It tells us about the magnitude of the association, or correlation, as well as the direction of the relationship.
It is applicable when the relationship is linear and its homoscedasticity.
Degree of correlation:
Perfect: If the value is near ± 1, then it said to be a perfect correlation
High degree: If the coefficient value lies between ± 0.50 and ± 1, then it is said to be a strong correlation.
Moderate degree: If the value lies between ± 0.30 and ± 0.49, then it is said to be a medium correlation.
Low degree: When the value lies below + .29, then it is said to be a small correlation.
No correlation: When the value is zero.

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

Answer : Scaling refers to putting the variable values into the same range.
When we have a lot of independent variables in a model, many of them might be on different scales which will lead a model with weird coefficients which might be difficult to interpret.
Standardized Scaling: The variables are scaled in such a way that their mean is zero and standard deviation is one.
MinMax Scaling (normalized 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.

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

Answer : If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. 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 (this also show an infinite VIF).

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

Answer : Quantile-Quantile (QQ) plot is to show if two data sets come from the same distribution. Plotting the first data set's quantiles along the x-axis and plotting the second data set's quantiles along the y-axis is how the plot is constructed. This helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

QQ plots is very useful to determine

1. If two populations are of the same distribution
2. If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.
3. Skewness of distribution