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. Fall Season has more demand followed by summer, winter and spring.

   ii. When the weather situation is clear, it shows better demand compared to mist and light rain conditions.

   iii. The demand for bikes were increased as the year moves from 2018 to 2019.

   iv. During the time of holidays, the demand got decreased as usage becomes rare as expected.

   v. The demand increases gradually from January to June and then decreases from June to December.

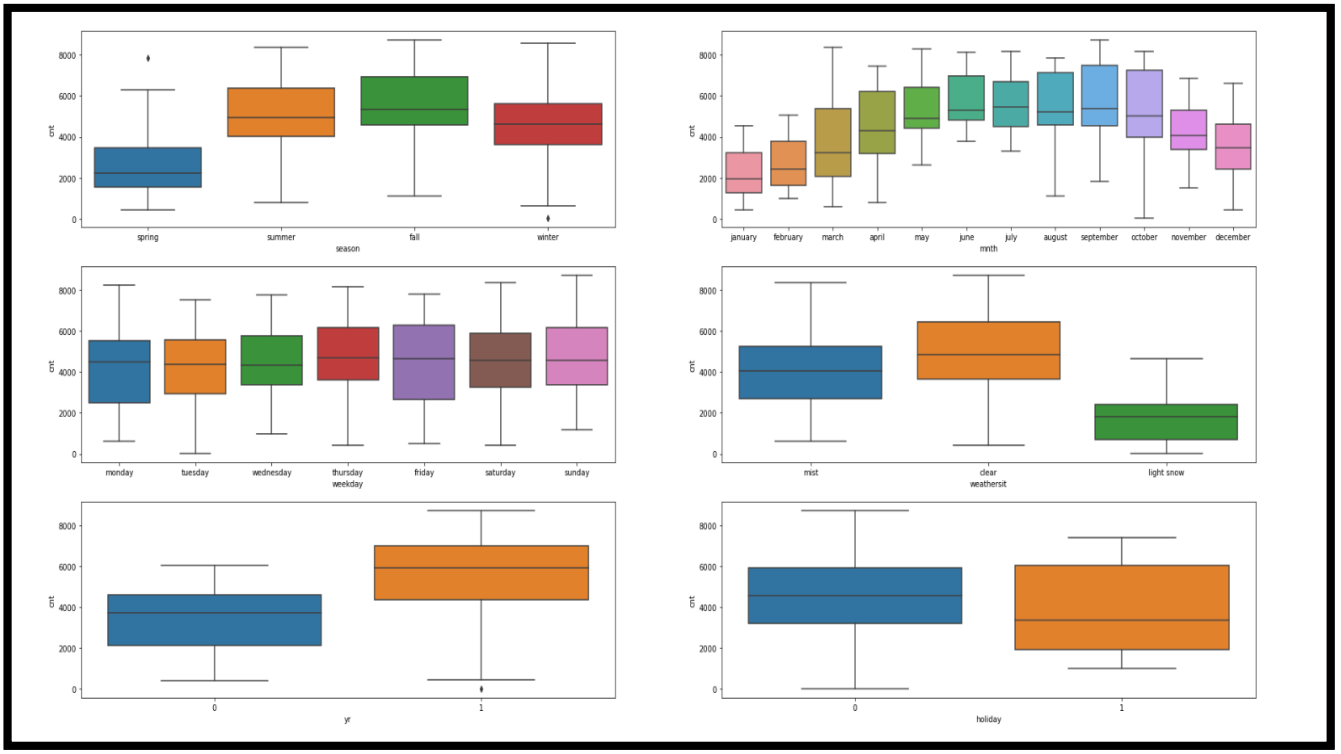

Figure 1: Target variables vs Categorical Variables

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

When we are converting the categorical variable with levels into dummy variables, the number of variables within the model increases. As these dummy variables carry binary levels it's easy for the model to identify the significance of the dropped column with the remaining columns available Hence we do not need that extra column. Typically, if there are n levels in a categorical variable then it's enough to create n-1 dummy variables.

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

The Registered variable has the highest correlation with the cnt variable before doing the feature scaling. (Considering all features).

The temp variable has the highest correlation with the cnt variable after doing the feature scaling. (Dropping multi-collinear features).

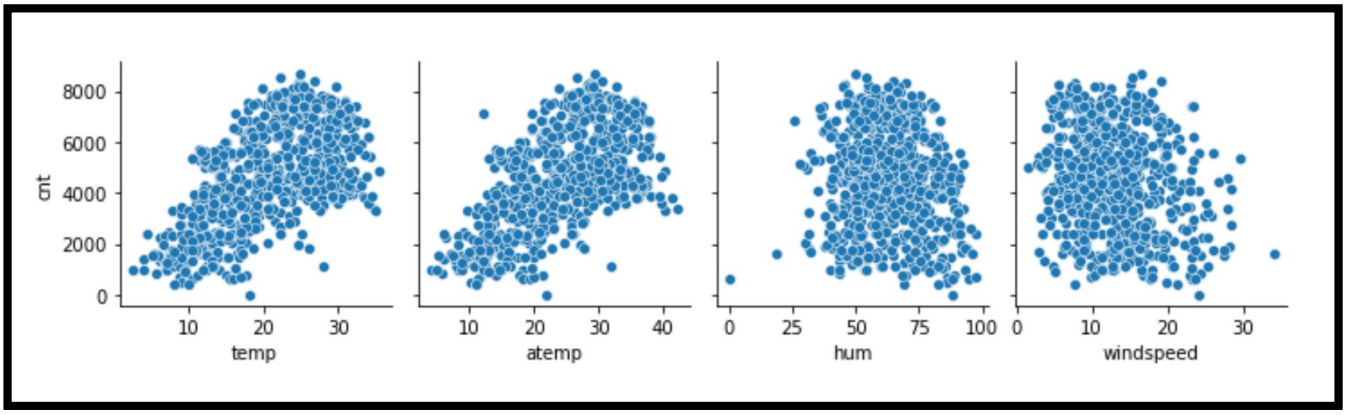

Figure 2: Target Variable vs Numerical Variables

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

i. Find the residuals or error terms, draw a distribution plot to check whether the residuals are distributed normally or not with the mean as zero – Error terms are normally distributed.

ii. Plot a scatter plot between residuals and independent variables to check for a particular pattern – Error terms are independent of each other.

iii. Plot a scatter plot between error terms vs target variable -Error terms have constant variance (homoscedasticity).

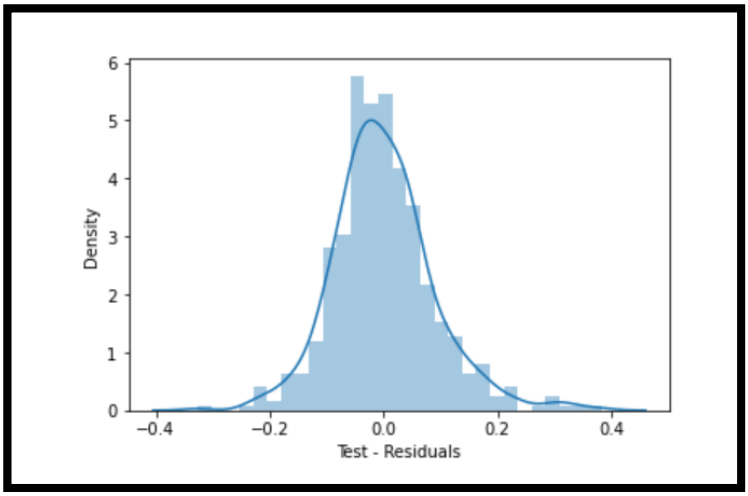

Figure 3: Residuals Distribution

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

Temp, weather light snow and year are the top 3 features contributing significantly towards the demand for shared bikes.

OLS Regression Results

Dep. Variable:	cnt	**R-squared:**	0.840
Model:	OLS	**Adj. R-squared:**	0.836
Method:	Least Squares	**F-statistic:**	237.0
Date:	Sun, 08 May 2022	**Prob (F-statistic):**	6.89e-190
Time:	18:23:54	**Log-Likelihood:**	505.58
No. Observations:	510	**AIC:**	-987.2
Df Residuals:	498	**BIC:**	-936.3
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2262	0.027	8.384	0.000	0.173	0.279
yr	0.2280	0.008	27.917	0.000	0.212	0.244
holiday	-0.0989	0.026	-3.844	0.000	-0.149	-0.048
temp	0.5977	0.023	26.494	0.000	0.553	0.642
hum	-0.1841	0.038	-4.876	0.000	-0.258	-0.110
windspeed	-0.1895	0.026	-7.351	0.000	-0.240	-0.139
season_summer	0.0815	0.011	7.580	0.000	0.060	0.103
season_winter	0.1347	0.011	12.732	0.000	0.114	0.155
mnth_july	-0.0478	0.018	-2.640	0.009	-0.083	-0.012
mnth_september	0.0962	0.016	5.957	0.000	0.064	0.128
weathersit_light snow	-0.2318	0.026	-8.757	0.000	-0.284	-0.180
weathersit_mist	-0.0502	0.011	-4.771	0.000	-0.071	-0.030

Omnibus:	52.291	**Durbin-Watson:**	2.069
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	104.838
Skew:	-0.600	**Prob(JB):**	1.72e-23
Kurtosis:	4.869	**Cond. No.**	18.4

Figure 4: Final Model Summary

General Subjective Questions

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

Linear regression is used when the response or target variable is continuous in nature and there is a linear relationship between response and feature variables. Based on a number of feature variables it is divided into 2 types. If the number of feature variables is one, it is called as Simple linear regression and if the number of feature variables is more than one, it is called as Multi Linear regression.

The best fit for the model is decided by the straight-line equation, such that the residual sum of squares is minimum. After estimating the coefficients m and c with satisfied R2 value, RSME, VIF and P values of features, the model is good to go for predicting the response variable.

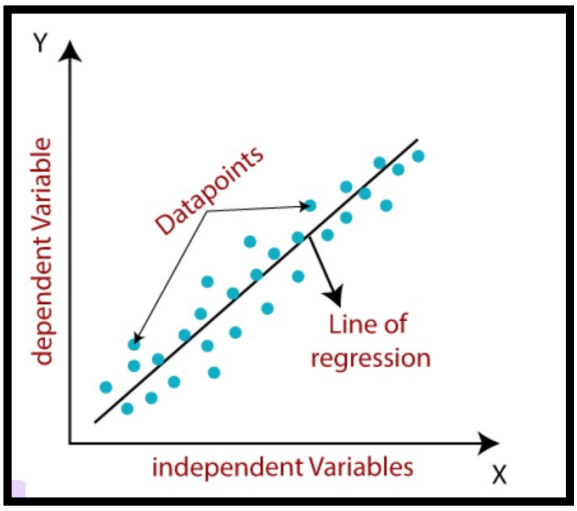

Figure 5: Linear Regression

Linear Regression is used to predict the continuous variable from the independent variables data points in such a way that its sum of residual squares (y actual – y predicted = residual) is minimum i.e., the cost function is minimum. The gradient Descent method is a useful algorithm that helps in minimizing the Cost function.

Assumptions:

- The relation between a dependent variable and an Independent variable is linear.
- Error terms are normally distributed.
- Error terms are independent of each other.
- Error terms have constant variance (homoscedasticity).
- No multicollinearity in predictor variables

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

The four datasets are almost identical concerning the simple descriptive statistics like mean yet visualized differently when plotted and following 4 different distributions.

Example:

S.no	x1	y1	y2	y3	y4
1	10	8.04	9.14	7.46	6.58
2	8	6.95	8.14	6.77	5.76
3	13	7.58	8.74	12.74	7.71
4	9	8.81	8.77	7.11	8.84
5	11	8.33	9.26	7.81	8.47
6	14	9.96	8.1	8.84	7.04
7	6	7.24	6.13	6.08	5.25
8	4	4.26	3.1	5.39	12.5
9	12	10.84	9.13	8.15	5.56
10	7	4.82	7.26	6.42	7.91
11	5	5.68	4.74	5.73	6.89
Mean	**9.0**	**7.5**	**7.5**	**7.5**	**7.5**
Variance	**10**	**3.75**	**3.75**	**3.75**	**3.75**
Standard Deviation	**3.16**	**1.94**	**1.94**	**1.94**	**1.94**

From the above data set, when we do descriptive statistics, the data sets are identical but visually they are different. Visualization as follows:

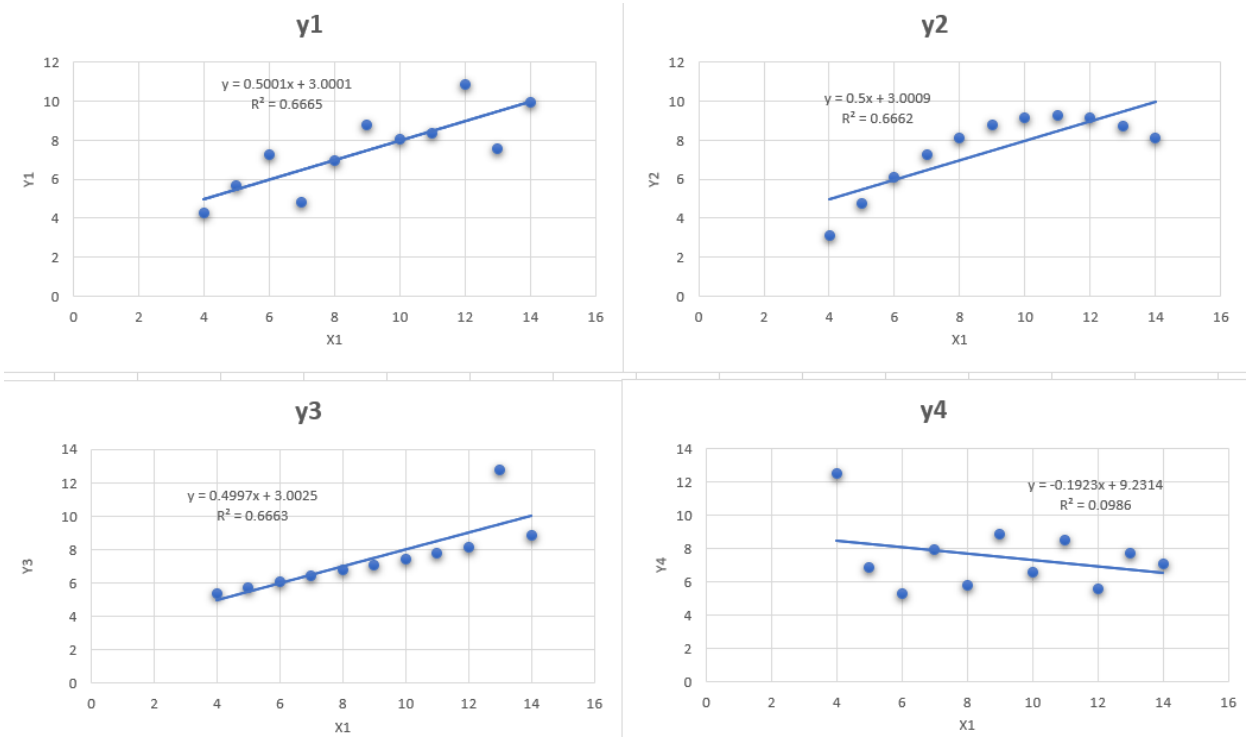

Figure 6: Anscombe Quartet Visuals

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

Pearson's R is Pearson Correlation Coefficient. It measures the association between two continuous variables. The r-value oscillates between -1 to +1. Depending upon the value we can infer the relation between the two variables.

Cases: (Generalized)

- If R is Positive and above 0.5, the variables are directly proportional and have a higher correlation.
- If R is Positive and lies between 0.5 to 0.3, the variables are directly proportional and have a medium correlation.
- If R is Positive and lies between 0.3 to 0, the variables are directly proportional and have a lower correlation.
- If R is Negative and above 0.5 in magnitude, the variables are Inversely proportional and have a higher correlation.
- If R is Positive and lies between 0.5 to 0.3 in magnitude, the variables are Inversely proportional and have a medium correlation.
- If R is Positive and lies between 0.3 to 0 in magnitude, the variables are Inversely proportional and have a lower correlation.
- If R is Zero, there is no relationship between variables.

The Formula to calculate R is

$$R = \frac{\sum_i^n ((X_i - Xmean)(Y - Xmean))}{\sqrt{\sum_i^n (X_{i-Xmean})^2} \sqrt{\sum_i^n (Y_{i-Ymean})^2}}$$

Where Xi is ith variable of X
Xmean is the mean of X
Yi is ith variable of Y
Ymean is the mean of Y

Technically, Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

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

Scaling is the process of bringing all the data points to the same scale with respect to the magnitude which helps in for easy interpretation. As Each predictor variable has different scales, it is important to scale the variables.

Normalization or Minmax scaling: It converts the entire data points to the scale of 0 to 1 with respect to their magnitude.

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

Standardized scaling: it converts the entire data points in such a way that its mean is zero and the standard deviation is 1.

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

The Advantage of Standardization is that it doesn't squeeze the data into a particular range. It is useful when the feature variables have more outliers.

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

VIF infinite means it has the R square value as 1 according to the formula. The R square is calculated within the predictor or independent variables i.e., making one of the independent variables as y_ variable and all the remaining independent variables as x_ variables. So, if there is any correlation between the predictor variables then obviously R square value becomes one which eventually leads to VIF as infinite. So, infinite VIF means R square is one amongst the predictor variables which reckons the multicollinearity between them.

$$VIF_i = \frac{1}{1 - R_i^2}$$

Where VIF i = ith predictor variable Variance Inflation factor

R i = R Square of ith Variable.

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

(3 marks)

Quantile - Quantile plots are graphical representations between the quantiles i.e., fractions of the data set.

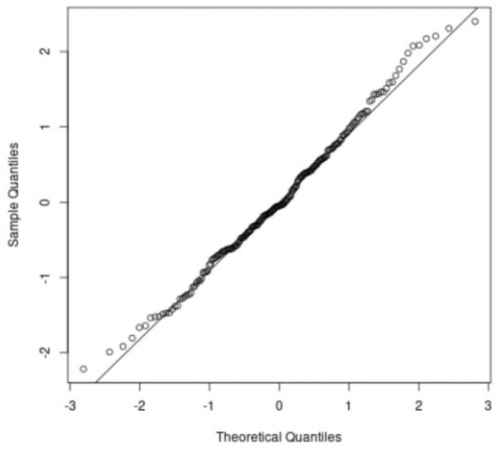

Figure 7: Normal Q-Q plot

The Purpose of the Q-Q plot is to whether two data sets are from the same distribution or not. By drawing the 45-degree line on the graph if the data points fall the reference line, then one can say they are from the same distribution.

A Q–Q plot is used to compare the shapes of distributions and whether skewness is similar or different in the two distributions.