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)

Ans:

- Fall season has highest demand for bike riding followed by the summer season
- The demand has increased in the year 2019 when compared to the year 2018
- Most demand for the bike is from the month of May to October with July being the highest month in demand.
- The demand is more fur working days i.e; on non-holidays (The data in column holiday is a derivative of the data in the column workingday
- All the seven days are having almost equal demand for bikes.
- The demand is highest when the weather is clean and cloudy and the least when it is raining or snowing

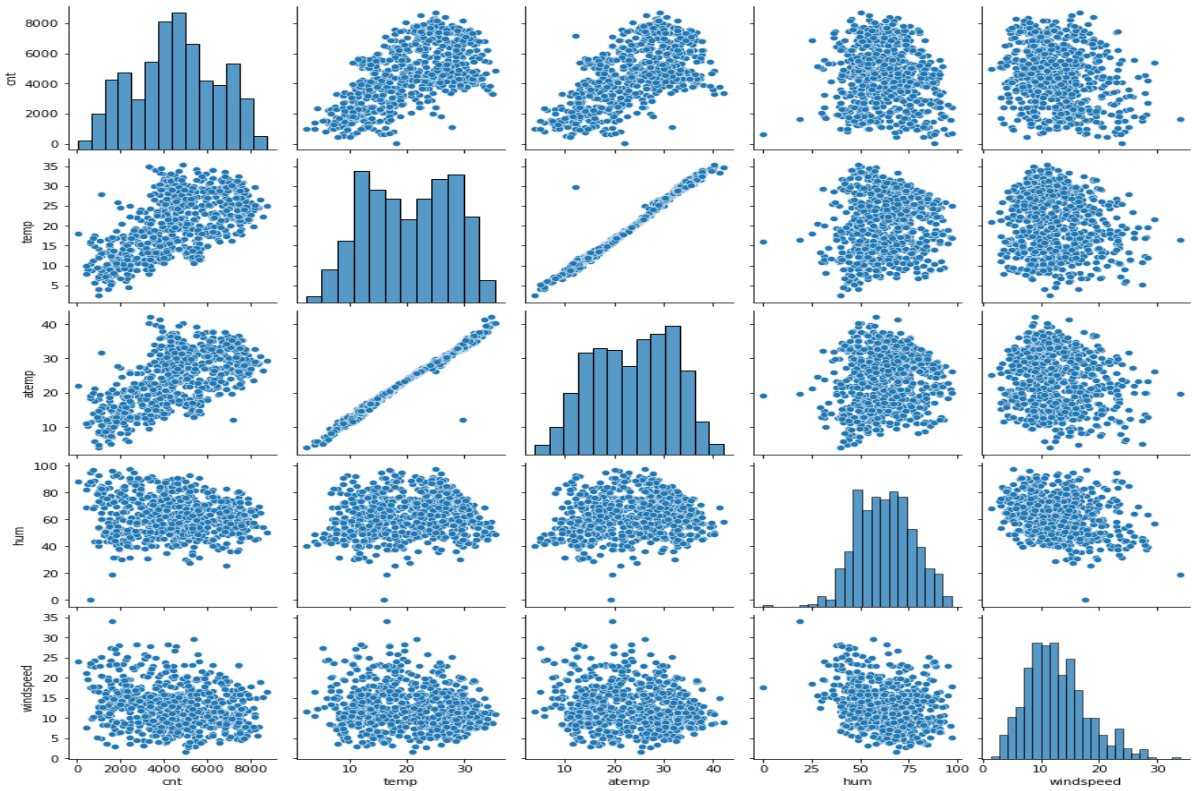
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Ans:

- We drop the first column in order to avoid correlation between the dummy variables
- This may adversely affect the model
- Also, we would like to avoid multicollinearity between the dummy variables

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

Ans:

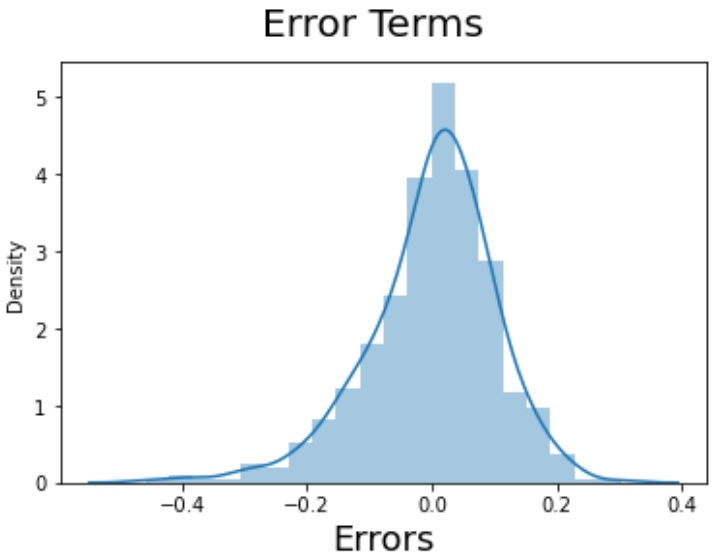


- Temp and atemp columns are highly correlated to the target variable cnt

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

Ans:

- We plot the distribution of Error to validate the assumptions
- We find that the errors, i.e the residuals are normally distributed with mean as zero.



-
-

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

Ans:

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.788
Model:	OLS	Adj. R-squared:	0.786
Method:	Least Squares	F-statistic:	312.3
Date:	Tue, 08 Feb 2022	Prob (F-statistic):	4.95e-166
Time:	07:31:51	Log-Likelihood:	434.47
No. Observations:	510	AIC:	-854.9
Df Residuals:	503	BIC:	-825.3
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0237	0.014	1.706	0.089	-0.004	0.051
yr	0.2327	0.009	25.088	0.000	0.215	0.251
temp	0.5959	0.022	27.490	0.000	0.553	0.638
season_2	0.0763	0.012	6.590	0.000	0.054	0.099
season_4	0.1336	0.012	11.540	0.000	0.111	0.156
mnth_9	0.0927	0.018	5.240	0.000	0.058	0.127
weathersit_3	-0.2658	0.027	-9.668	0.000	-0.320	-0.212

Omnibus:	64.668	Durbin-Watson:	1.970
Prob(Omnibus):	0.000	Jarque-Bera (JB):	111.222

```
Skew:                          -0.783   Prob(JB):                 7.06e-25
Kurtosis:                       4.668   Cond. No.                     7.91
==============================================================================
```

- The variable temp with coefficient 0.5959
- The variable yr with coefficient 0.2327
- The variable weathersir_3 (Light Snow and Rain) with coefficient -0.2658

General Subjective Questions

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

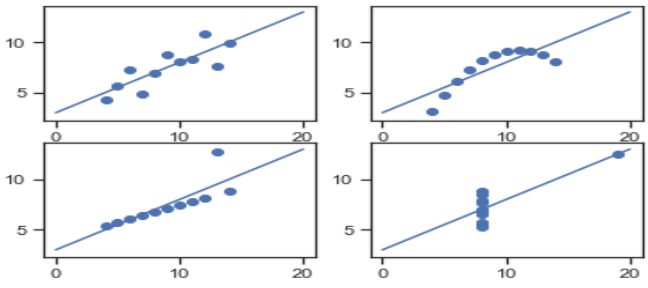
Ans:

- Linear regression is a machine learning algorithm based on supervised learning.
- Regression models a target prediction value based on independent variables
- Linear Regression finds out the linear relationship between the x (input) and y (output). Hence the name Linear Regression
- Regression is broadly divided into Simple Linear Regression (SLR) and Multiple Linear Regression (MLR)
- SLR is used when the dependent variable is predicted using only on independent variable
- The regression line is the best fit line for the model which is given as:
    - Y = mx + c (SLR)
    - $\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$ .....

    Where b_0 or c is the intercept and m, b_1, b_2, etc are coefficients

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

Ans:

- Anscombe's quartet comprises of 4 data sets that have nearly identical simple statistical properties but appear very different when graphed.
- It was developed by statistician Francis Anscombe in 1973 to emphasis on the importance of graphing data before analysing it and the effect of outliers on statistical properties
- 

Pic Courtesy: https://medium.com/analytics-vidhya/anscombes-quartet-an-importance-of-data-visualization-856b3d1bd403

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

Ans:

- It is the ratio between the covariance of two variables and the product of their standard deviations.
- It is essentially a normalized measurement of the covariance; hence the result always has a value between −1 and 1.
- It was developed by Karl Pearson from a related idea introduced by Francis Galton in the 1880s, and for which the mathematical formula was derived and published by Auguste Bravais in 1844.
- It shows a linear relationship between two sets of data where
    - r = 1 means data is perfectly linear with positive slope
    - r = -1 means data is perfectly linear with negative slope
    - r = 0 means there is no linear association

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

Ans:

- When we have a lot of independent variables in a model, a lot of them might be on different scales which will lead to a model with coefficients that might be difficult to interpret. Hence we need to scale features for
    - Ease of interpretation
    - Faster convergence for Gradient Descent methods
- There are two methods for scaling
    - Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$$

    - MinMax 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.

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

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

Ans:

- VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity
- 〖VIF〗_1=1/(1-R_1^2) ; R-1 is the R Square value of the independent variable for which we are checking the correlation with other independent variables
- If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

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

Ans:

- A Q-Q plot is a plot of quantiles of the first data set against the quantiles of the second data set.
- It is used to compare the shapes of distributions
- It is scatter plot of two set of quantiles against one another.
- If both the quantiles come from the same distribution, the points will be formed in a roughly straight line
- It is used to find out
    - Whether the data sets have come from the population with common distribution
    - Whether the data sets have common location and scale
    - Whether the data sets have similar distributional shapes
    - Whether the data sets have similar tail behaviour