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:

Firstly, we have totally **seven** categorical variables in our dataset after data cleaning. When compare these variables with our target variable – **count.** I can see,

- Season – **Fall** and **Summer** seasons have more rides while comparing the other two seasons.
  - ✓ Fall > Summer > Winter > Spring.
- Month - From **May** to **Sep**, we have more no. of rides while comparing start and end of the year.
- Weekday - Couldn't see much of a pattern. Data is almost equally scattered among the target variable.
- Weather sit – Out of given four levels in data dictionary. We have only three of them in our dataset. Among them, the correct pattern has been followed.
  - ✓ **Clear** weather has more count than **Mist** and
  - ✓ **Mist** has more count than **Light rain/snow**
  - ✓ Clear > Mist > Light rain/snow
- Year – Clearly **2019** have higher count when comparing to **2018**.
  - ✓ 2019 > 2018
- Holiday – In holidays we have lesser rides. So, we could say the rides mostly used by students or working personals.
- Working day – Same as holiday, in working day we have more rides. But not very much.

And when we look at the correlation chart, we could see variable **year** is the highly correlated variable with the target variable when comparing the other categorical variables.

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

Ans:

Let's say we have a variable with three levels. If we need to create dummy variables for this. i.e., changing this variable into multiple variables so that it could have only binary values.

For our case we would have,

Level 1	Level 2	Level 3
1	0	0
0	1	0
0	0	1

Let's create another one with just two of them,

Level 1	Level 2
1	0
0	1
0	0

Now, with these two levels itself, we can predict the third one.

- ❖ In first and second row there are no possibilities of '1' in third level since we would have one in anyone of the level at a time. So, since we have 1 in first two rows already, third level will definitely have '0'.
- ❖ In the third row, in level one and two we have zero, we have only one remaining level so it will take '1' for sure.

So clearly, we can predict the third level with the two levels we have. So, we don't need third one.

That's what **drop_first=True** is doing here. It will drop the first level of the variable and will give n-1 levels of n levels.

If we didn't use **drop_first=True**, we will get affected by multicollinearity. As the two levels of the variable fully explains the third one, we will get high VIF. Hence, we will drop it there.

Instead, we are using **drop_first=True** and dropping one of the levels while creating dummy variables itself.

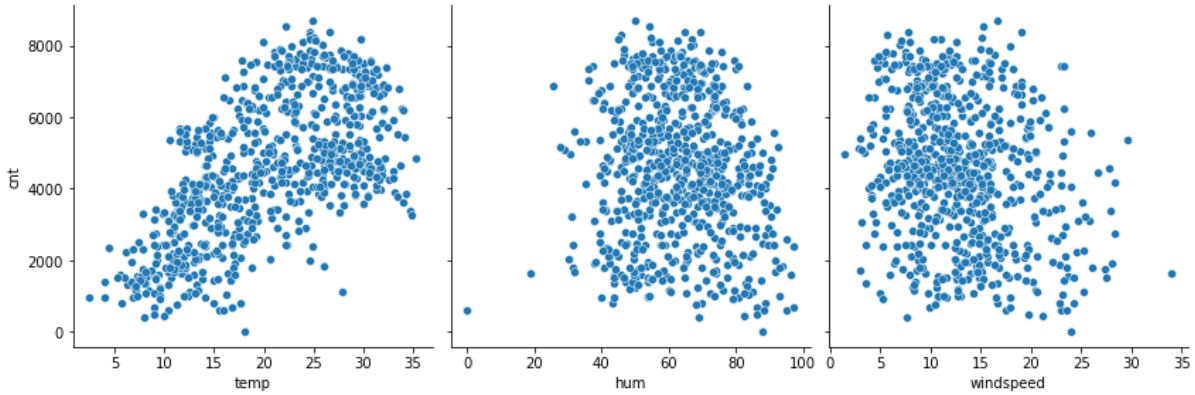
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans:

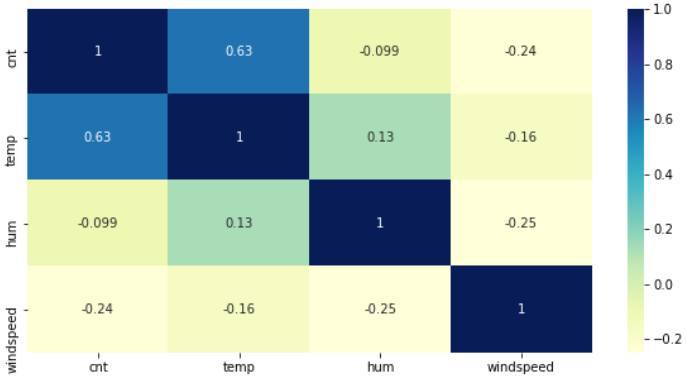
Firstly, we have totally **four** numerical variables in our dataset after data cleaning.

Note: When we see the correlation of temp and atemp. It is **99%** correlated. So, I dropped one of them i.e., atemp.

So now, from the three numerical variables, **temp** is highly correlated with the target variable - **Cnt**.



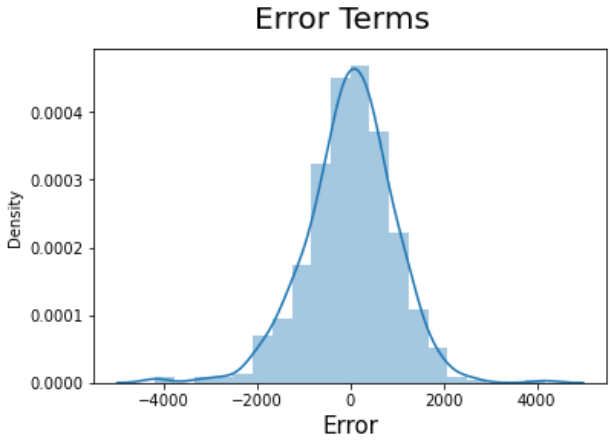
Temp and **Cnt** are **63**% correlated.



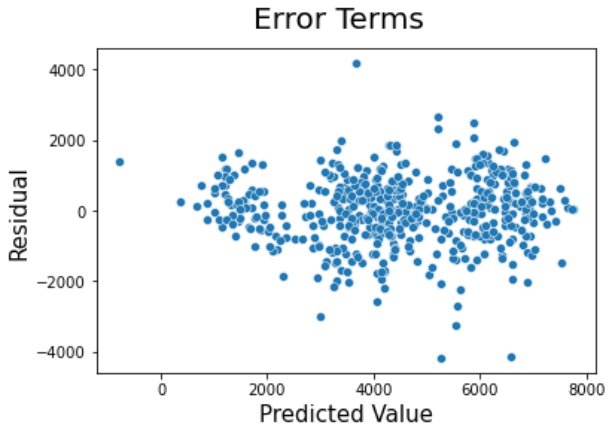
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

We have three major assumptions,

1. **Normality of Error** - Error values are normally distributed with mean 0.
   - So plotted a distribution plot for the residuals (y train – y train pred).
   - And yes, we did get the normally distributed error values with mean 0.



2. **Homoscedasticity** – The probability distribution of errors has constant variance.
   - We can see that most of the values equally divided from zero and have same distance. Hence it follows Homoscedasticity.
3. **Independence of Error -** Error values are statistically independent.
   - We couldn't see much of a pattern in the error values. So, it is statistically independent.



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:

With the final model, the top 3 features contributing significantly towards explaining the demand of the shared bikes are

- **Season - spring** - with the coefficient of **-2596.8759**
- **Weather sit - light rain/snow** - with the coefficient of **-2585.2421**
- **Year** – with the coefficient of **2151.7311**

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.764
Model:                            OLS   Adj. R-squared:                  0.759
Method:                 Least Squares   F-statistic:                     179.4
Date:                Tue, 31 Aug 2021   Prob (F-statistic):          2.44e-150
Time:                        17:05:46   Log-Likelihood:                -4219.6
No. Observations:                 510   AIC:                             8459.
Df Residuals:                     500   BIC:                             8502.
Df Model:                           9
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                         5111.5319    128.742     39.704      0.000    4858.591    5364.473
yr                            2151.7311     85.144     25.272      0.000    1984.447    2319.016
windspeed                    -1498.5568    260.841     -5.745      0.000   -2011.037    -986.077
season_spring                -2596.8759    129.313    -20.082      0.000   -2850.940   -2342.812
season_summer                 -362.7919    128.047     -2.833      0.005    -614.368    -111.216
season_winter                 -653.8492    123.187     -5.308      0.000    -895.877    -411.822
mnth_Sep                       611.8474    168.319      3.635      0.000     281.147     942.548
weekday_Tu                    -396.7705    121.560     -3.264      0.001    -635.601    -157.940
weathersit_Mist               -788.0897     90.839     -8.676      0.000    -966.563    -609.616
weathersit_light rain/snow   -2585.2421    257.207    -10.051      0.000   -3090.582   -2079.903
==============================================================================
Omnibus:                       37.189   Durbin-Watson:                   2.003
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               88.022
Skew:                          -0.379   Prob(JB):                     7.70e-20
Kurtosis:                       4.889   Cond. No.                         8.74
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

General Subjective Questions

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

Ans:

- Linear regression shows relationship between continuous variables.
- In linear regression, we find a best fitted line among the data points of the target variable. Here best fitted line is nothing but the distance between the line(predicted) and each data point(actual) also called as residuals, is minimum.
- It shows the linear relationship between the independent variable and the dependent (target) variable.
- If there is only one independent variable then it is simple linear regression.
- If there are more than one independent variable then it is multiple linear regression.
- We have our straight-line equation

  $$Y = mx + c$$

  Here, we will use it has

  $$y = \beta_0 + \beta_1 X$$

where,

y – dependent variable

X – independent variable

β_0 – Intercept

β_1 – Slope/ Co-efficient

By finding the best value of β_0 & β_1 in a way the residuals are minimum. we will get a best fitted line. To do this, we use OLS (Ordinary Least Squared) and get the best fitted line.

In simple linear regression, it is simple. By using OLS directly we will get our results.

But in multiple regression, we need to find which variable speaks well about the target variable. And need to keep those variables in our model.

We have some factors here which helps us to get the best result,

- **R-squared** – Which tells us how well the model explains the target variable.
- **Adjusted R-squared** – which is R-squared with some penalty to remove redundant variables in the model.
- **p-value** – Whether a particular variable is significant or not.
- **F-stat** – Whole model is significant or not.
- **VIF** – To avoid multicollinearity.
- Residual Analysis – Assumptions of linear regression.
  - ✓ **Normality of Error** - Error values are normally distributed with mean 0.
  - ✓ **Homoscedasticity** – The probability distribution of errors has constant variance.
  - ✓ **Independence of Error -** Error values are statistically independent.

When the r2-score of train and test data is closer. We can say our linear regression model is a good fit. And it will predict the future/new data very well.

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

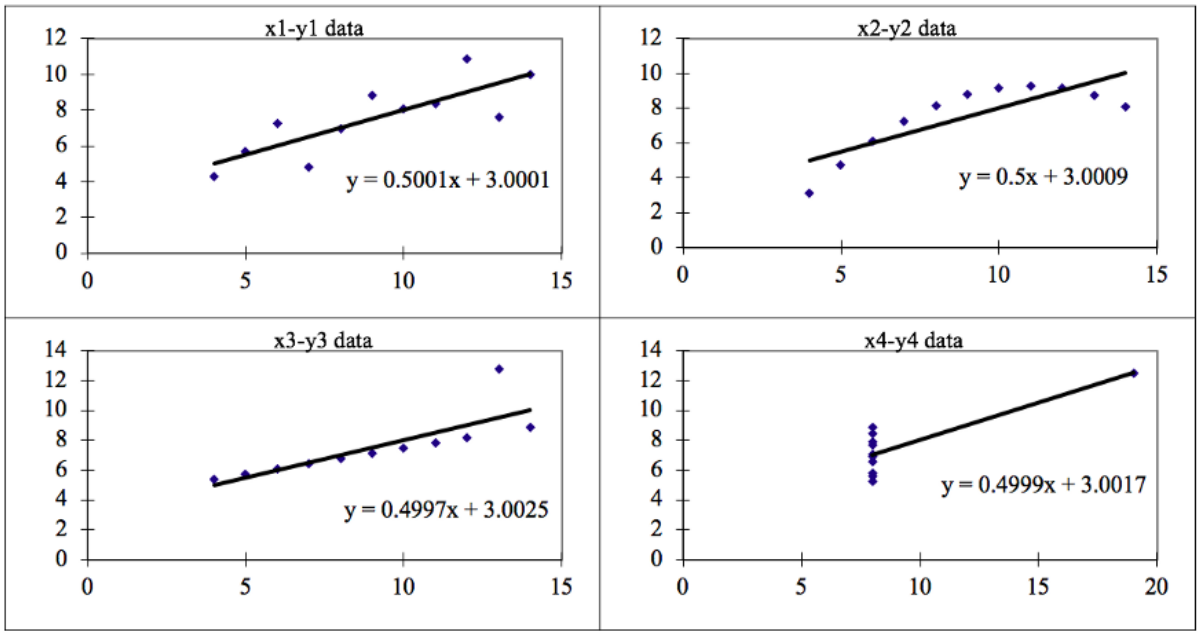
Ans:

- Anscombe's quartet tells us how important is data visualization.
- Anscombe's quartet contains four data sets which will have very identical descriptive statistics, but have very different distributions and appear very different when plotted using scatter plot.
- It was constructed by Francis Anscombe in 1973 to tell how important to do data visualization before model building.

Given below Anscombe's data,

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

We can see mean, standard deviation and r are identical. But when we see the distributions,



We can clearly see we have four very different graphs here.

- Data Set 1: Fits the linear regression model very well.
- Data Set 2: Won't fit for the linear regression model since it is non-linear.
- Data Set 3: Have outlier which cannot be handled by linear regression model.
- Data Set 4: Have outlier which cannot be handled by linear regression model.

From this we can clearly see the importance of the data visualization before model building. As it fooled us with the statistics. And only after plotting we got the clear and correct picture of the data. So, we must do data visualization before model building.

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

Ans:

- Pearson's R also called as Pearson Correlation Coefficient tells us the strength of the linear relationship between two variables. Denoted by **r**.
- Pearson Product-Moment Correlation (PPMC in short) tries to draw a best fitted line through the data of two variables. Pearson Correlation Coefficient, r, indicates how well the data points are fit this best fitted line.
- r can range from -1 to 1.
  - If the value is greater than 0, it indicates, if one variable increases another variable also increase i.e., Positive Correlation.
  - If the value is lesser than 0, it indicates, if one variable increases another variable will decrease i.e., Negative Correlation.
  - If the value equals to 0, then there is no relation between the two variables.
- If the association between two variables is stronger, the r value will be closer to -1 or 1 depends on if it is negative or positive association.
- If the association is weaker, the r value will be closer to 0. Which means greater the variation around the best fitted line.
- We can say the strength of the association as
  - Small – (0.1 to 0.3) or (-0.1 to -0.3)

- Medium – (0.3 to 0.5) or (-0.3 to -0.5)
- Large – (0.5 to 1.0) or (-0.5 to -1.0)

- Slope is not equal to r. The r just says the variation between the data points and the best fitted line. R increases, variation decreases.
  - Let's say we have r = 1, It doesn't mean if one unit of A variable increase, there will be a unit increase in the B variable. It just says We have no variation between the data points and the best fitted line.

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

Ans:

- ✓ Scaling is nothing but a technique used to standardize an independent variable in a data set to a fixed range. We will do this after data split of train and test.
- ✓ In linear regression, the range of coefficients may vary widely among the independent variables if we don't perform scaling. In other words, the data we receive may vary highly in range, magnitudes or units. The linear regression will take only magnitude and not units, So we may end up in incorrect modelling.
- ✓ Since linear regression is very sensitive for the range of data, by doing scaling, we will get better results.

S.NO.	Normalized Scaling	Standardized Scaling
1.	For scaling, Minimum and Maximum values are used.	For scaling, Mean and standard deviation is used.
2.	Scales values between (0,1) or (-1,1).	There is no bound here.
3.	Useful if features are of different scales.	Useful if we need to ensure zero mean and unit standard deviation.
4.	Affected by outliers	Less affected by outliers.
5.	We use MinMaxScaler for Normalization.	We use standardScaler for Standardization.
6.	Also called as Scaling Normalization.	Also called as Z-score Normalization.

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

Ans:

- ➢ The VIF will be equal to infinity if there is a perfect correlation between independent variables.
- ➢ The Infinite VIF indicates the corresponding variable is completely explained by the other independent variables.
- ➢ So, if there is an independent variable completely explained by the other independent variables. Then the R-Square will become exactly 1.
- ➢ We know VIF = $1/(1-R^2)$. If $R^2 = 1$ then the equation becomes 1/(1-1) => 1/0 => Infinity.

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 used to compare the shapes of the distribution, providing a graphical view of how properties of location, scale and skewness are similar or different in the two distributions.

- It is used to check, If two data sets
  - ✓ Come from population of common distribution.
  - ✓ Have common location and scale.
  - ✓ Have similar distributional shapes.
  - ✓ Have similar tail behaviour.
- A Q-Q plot is a plot of quantiles of two data sets
  - ✓ Similar Distribution: If all points of quantiles lie on or close to the straight line of 45 degrees.
  - ✓ X – values > Y – Values – If Y quantiles are lesser than the X quantiles.
  - ✓ X – values < Y – Values – If X quantiles are lesser than the Y quantiles.
  - ✓ Different Distribution: If all the points are lie away from the straight line of 45 degrees.