Linear Regression Assignment Subjective questions

Assignment based subjective questions

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

Ans:

- Categorical variables strongly affect the dependent variable (in this case the demand for bikes) here as they are among the top contributors to the demand.
- In fact, the top contributor to the demand is yr (0 – 2018, 1- 2019)
- Top 3 contributors have two categorical variables i.e. yr and season.
- The other categorical variables, even though having a negative coefficient, do increase the overall accuracy of the model as we saw in the assignment.

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

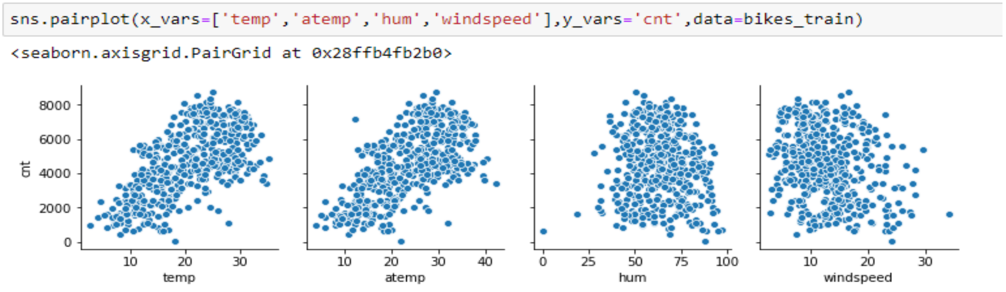
Ans:

- pandas.get_dummies() has an argument drop_first which can be set to True or False, depending on if we want to drop the first level and get p – 1 dummies from p values of a categorical variable
- We can drop the first level as we can conclude if any row will have the value of first level, based on the 0,1 values of other levels of dummies
- Eg:- For the gender column having 2 values i.e. Male and Female, we can just get 1 dummy variable i.e. Female, having values 0 and 1 as if Female = 0, then it means the person is Male.

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

Ans:

From the pair plots seen of the numeric variables and the target variable, the variable atemp seems to have the highest correlation with the target variable



```
sns.pairplot(x_vars=['temp','atemp','hum','windspeed'],y_vars='cnt',data=bikes_train)
```

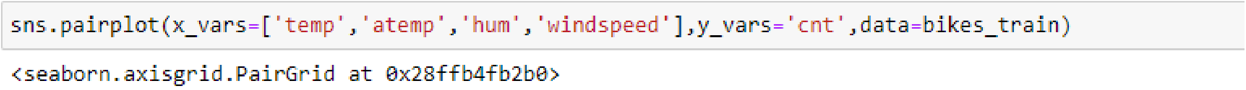
```
<seaborn.axisgrid.PairGrid at 0x28ffb4fb2b0>
```

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

Ans:

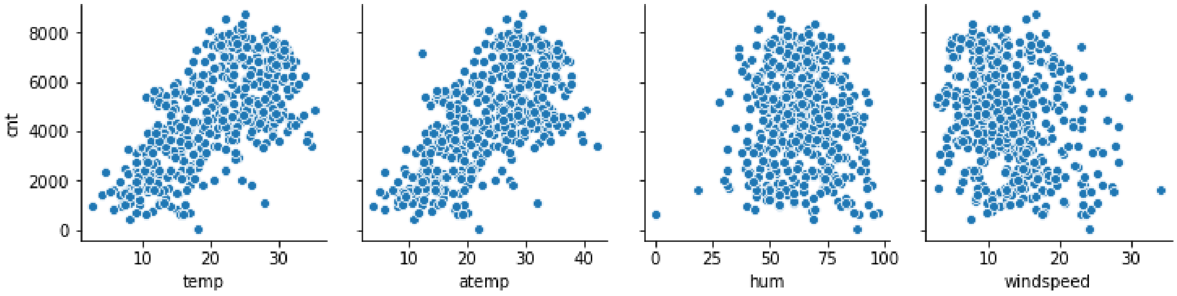
Linear Regression has the following assumptions:-

1. Linear relationship between the features and target variable
   From the pair plot between numeric variables and the target variable, we can see that linearity exists.

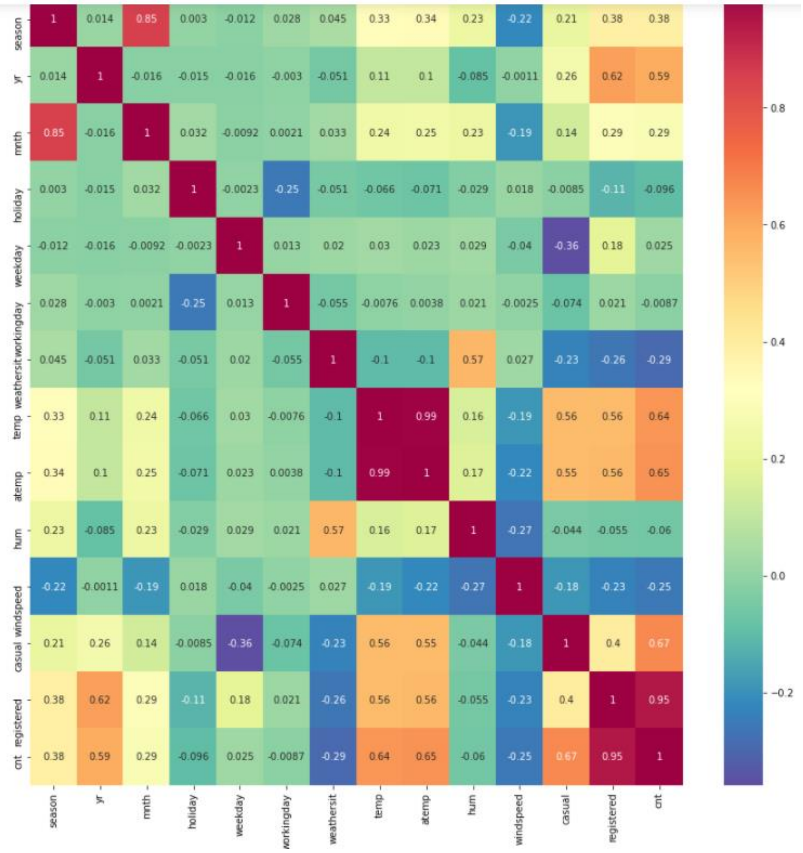


```
sns.pairplot(x_vars=['temp','atemp','hum','windspeed'],y_vars='cnt',data=bikes_train)
```

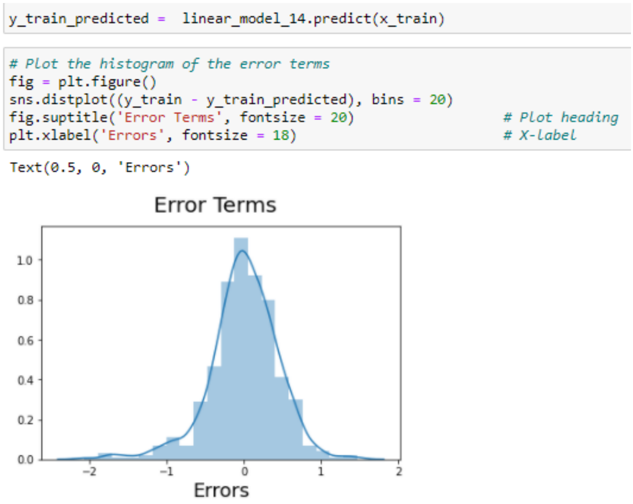
```
<seaborn.axisgrid.PairGrid at 0x28ffb4fb2b0>
```



2. Little or no multi collinearity between the features
   Multi collinearity can be checked by heatmap between the features



3. Error terms are distributed normally with mean zero



```
y_train_predicted =  linear_model_14.predict(x_train)
```

```
# Plot the histogram of the error terms
fig = plt.figure()
sns.distplot((y_train - y_train_predicted), bins = 20)
fig.suptitle('Error Terms', fontsize = 20)                  # Plot heading
plt.xlabel('Errors', fontsize = 18)                         # X-label
```

```
Text(0.5, 0, 'Errors')
```

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

Ans:

The top 3 features that are contributing significantly towards the target variable are **yr, atemp** and **season**.

General Subjective Questions

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

Ans: Linear Regression algorithm generally has the below steps

1. Reading and Understanding data
    a. Importing the data using pandas library
    b. We use libraries like read_csv, read_excel to read the data set from csv/xlsx files into a DataFrame
    c. Understanding the structure of the data
    d. Check the shape, info of the DataFrame and describing it to understand various stats about the DataFrame
2. Visualizing the data
    a. We use various plotting libraries to construct plots between various variables
    b. We use pair plots to plot numeric variables
    c. We use boxplots to plot the categorical variables
    d. We use heatmap to find out the correlation between different variables
3. Performing Linear Regression
    a. We use different approaches to construct the model in python, like sklearn and statsmodel.api

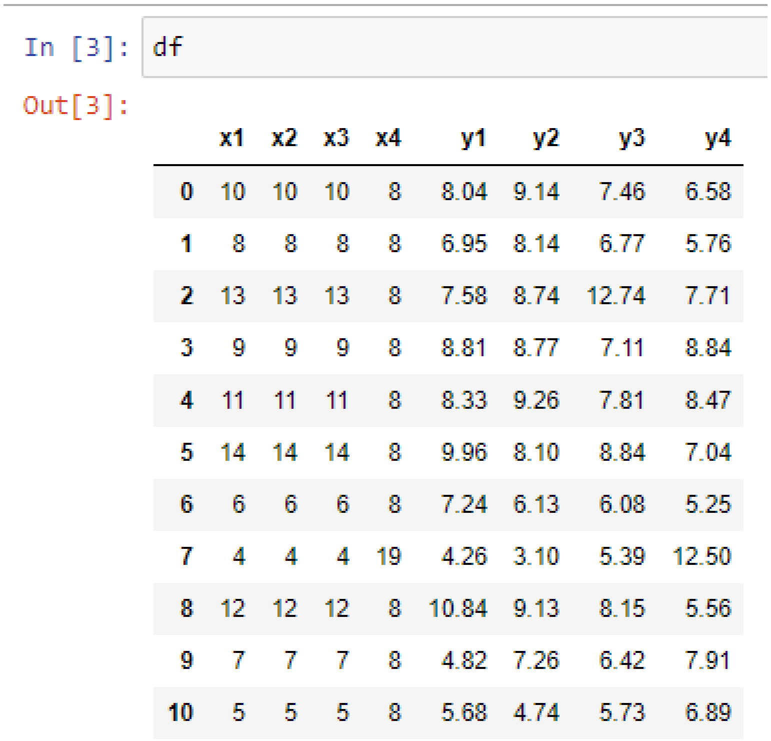
b. Once the model is created, we check if all the considered variables are significant by checking their respective p values and check the model accuracy by checking the R square and adjusted R Square

4. Residual analysis
   a. We check if the error terms are distributed normally with mean being zero and confirm if there is no clear relation between the training set and the residuals
5. Predictions on the data set
   a. We now make predictions on the data set using the model we constructed and then calculate various metrics like R2 score, Root Mean Square Error etc to understand the model accuracy.

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

Ans:

- Anscombe's quartet comprise of four data sets that have identical simple statistical properties but appear very different when graphed.
- Each data set consist of 11 (x,y) points.
- They were constructed in 1973 by statistician Francis Anscombe to demonstrate the importance of graphing data before realizing it and the effect of outliers on statistical properties
- Once Anscombe found 4 sets of 11 data points in his dream and requested the council as his last wish to plot those points. Following are the sets of 11 data points

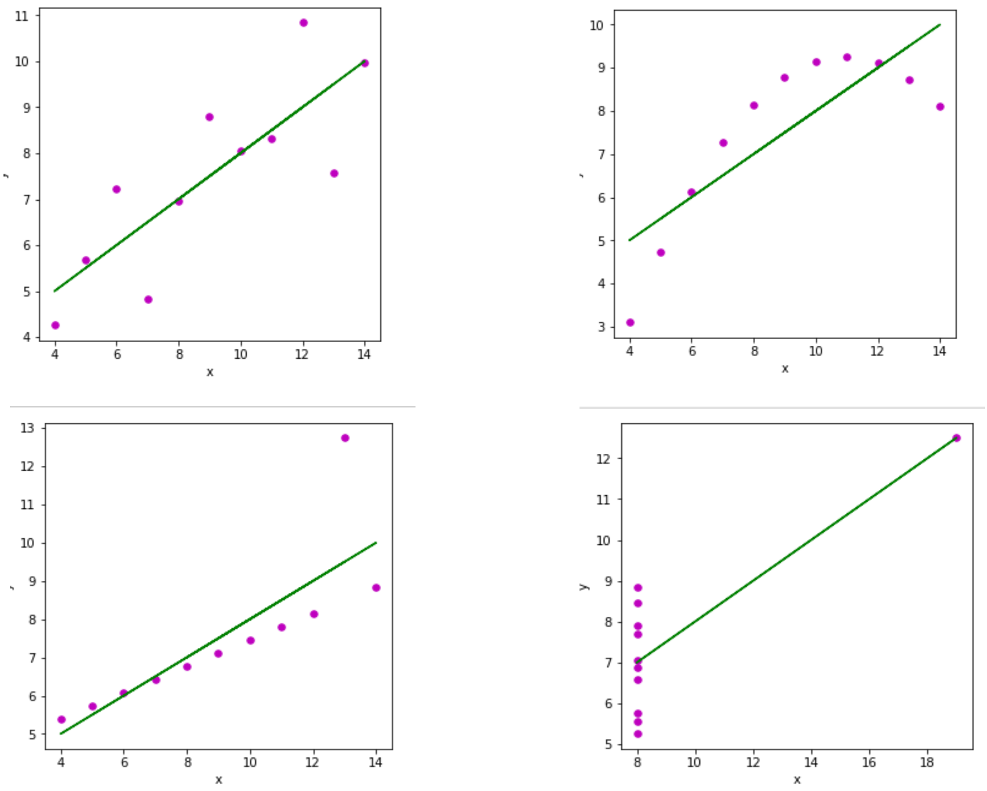


```
In [3]: df
```

Out[3]:

	x1	x2	x3	x4	y1	y2	y3	y4
0	10	10	10	8	8.04	9.14	7.46	6.58
1	8	8	8	8	6.95	8.14	6.77	5.76
2	13	13	13	8	7.58	8.74	12.74	7.71
3	9	9	9	8	8.81	8.77	7.11	8.84
4	11	11	11	8	8.33	9.26	7.81	8.47
5	14	14	14	8	9.96	8.10	8.84	7.04
6	6	6	6	8	7.24	6.13	6.08	5.25
7	4	4	4	19	4.26	3.10	5.39	12.50
8	12	12	12	8	10.84	9.13	8.15	5.56
9	7	7	7	8	4.82	7.26	6.42	7.91
10	5	5	5	8	5.68	4.74	5.73	6.89

- When graphed, the 4 data sets look like below



- Explanation of the above graphs
  - In the first one, there is almost a linear relationship between x and y
  - In the second one, there is a non-linear relationship between x and y
  - In the third one, there is a perfect linear relationship between x and y
  - In the fourth one, one high point exists which produces a high correlation coefficient
- Application
  - These quartets are still often used to illustrate the importance of looking at a data set graphically before starting to analyze according to a particular type of relationship and the fact that basic statistics are sometimes inadequate to describe data sets.

What is Pearson's R? (3 marks)

Ans:

- Pearson's R is a measure of linear correlation between two sets of data.
  It is the co-variance of two variables, divided by the product of their standard deviation.
- It was developed by Karl Pearson.
- Pearson's R is represented by the Greek letter ρ (rho).
- Given a pair of random variables (X,Y), the formula for ρ is:

$$\rho X, Y = \frac{\mathrm{cov}(X, Y)}{\sigma X. \sigma Y}$$

cov -> covariance

σ_X -> standard deviation of X

σ_Y -> standard deviation of Y

- Mathematical properties
  - The absolute values of Pearson's R are between 0 and 1
  - Correlations equal to +1 or -1 correspond to data points lying exactly on the line
  - It is symmetric i.e. corr(X,Y) = corr(Y,X)
  - A value of 1 implies that a linear equation describes the relationship between X and Y perfectly i.e. Y increases as X increases whereas a value of -1 implies that Y decreases as X increases.

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

Ans:

- Scaling is a step of pre-processing the data, which is applied to independent variables to normalize their values within a particular range. It also helps to speed up operations.
- The data that we collected will have values varying across a wide range of magnitude and units. If there is no scaling done, all values of the variables will be in different ranges and we will face issues analyzing the data.
- Eg:
  - A housing data set contains columns like price, area, num_bedrooms, num_bathrooms etc.
  - The price column has values in a different range as compared to the area column as compared to rest of the columns like bedrooms, bathrooms etc.
  - If the data is used as is to create a model, we might see smaller coefficients for price and area and large coefficients for bedrooms and bathrooms which might lead us to believe that certain variables contribute less/more to the target variable which might not be the case in reality.
  - That why we bring all the data to a certain scale to easy the model's processing as well as to correctly understand each variable's contribution to the target variable.
- Types of Scaling
  - Normalization/ min-max scaling
    - It brings all the data in the range of 0 and 1. sklearn. preprocessing.MinMaxScaler helps to implement this in python

$$Min - \max scaling: x = \frac{x - \min(x)}{\max(x) - \min\,(x)}$$

  - Standardization scaling
    - Standardization replaces the values by their Z scores. It brings all the data into a standard normal distribution which has mean (μ) zero and standard deviation (σ) one.

$$Standardization: x = \frac{x - \mu(x)}{\sigma(x)}$$

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

Ans:

- Multicollinearity is a phenomenon where independent variables are collinear. Collinearity refers to a linear relationship between two variables.
- VIF(Variance Inflation Factor) is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity.
- To determine VIF, we fit a model between the independent variables and refer the following table to understand the range of values in which VIF lies

VIF	**Conclusion**
1	No multicollinearity
4 - 5	Moderate
10	Severe

- Theoretically, we eliminate the variables having VIF values as high as above 5.
  In the industry, we consider VIF values less than 2 to be a good measure and try to discard anything above that.
- If there is a perfect correlation, then VIF will be infinity. So, if we are seeing that certain variables have VIF = infinity then there is a strong multicollinearity and we need to take care of that before proceeding further

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

Ans:

- Q Q Plots (Quantile – Quantile plots) are plots of two quantiles against each other. The range of values of a certain variable can be divided into quarters or quantiles (25%, 50%, 75%) .
- The purpose of Q Q plots is to find out if two data sets of data come from the same distribution.
- A 45-degree angle line is plotted on the Q Q plot: if the two data sets come from a common distribution, the points will fall on that reference line