Assignment-based Subjective Questions

- Palak Kaur –

Q1**. <u>From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)</u>**

Ans1. The categorical variables are insightful to understand the affect of various seasons, weekdays, working day, weather, months and year affect the 'cnt' value. We inferred the following –

- People tend to hire bikes in season_2 i.e. summer and season_4 i.e. winter. Summer tends to increase 'cnt' by 920 times, and 1055.8 times in winter season.
- The coefficient of yr_1 (i.e. year 2019) indicates that year 2019 year contributed 2025.16x as compared to year 2018.
- Months August, September and October have large positive impact on the values of cnt. Mnth_8 (August) has coefficient 552.4, Mnth9 (September) has coefficient 1057.1, and Mnth_10 (October) has coefficient of 511.0, indicating large increase of cnt in these months, largest in September month.
- High value of weekday_6 and workingday_1 coefficients (527.3 and 392.6 respectively) suggests that people are more likely to bike on weekdays and Saturday.
- Lastly, weathersit_2 (-437.2888) and weathersit_3 (-1861.68) explain the people are less likely to bike in these conditions : Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist and Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

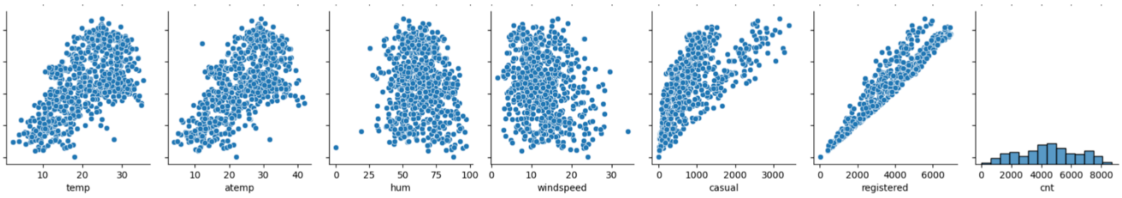
<u>Q2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)</u>

Ans2. It is important to use drop_first=True to avoid redundancy and multi-collinearity in data. It promotes simplified model with less complexity i.e. easy to model and predict.

For example, a dataframe has a categorical variables say, no_of_kids with values of 0,1,2 and 3. We can create 4 dummy variables as follows : 1. Kids_0, 2. Kids_1 3.Kids_3, Kids_4. We could reduce redundancy by dropping first element, thus if all other three values of Kids_1, Kids_2, Kids_3 have value 0, then it means it is Kids_0.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

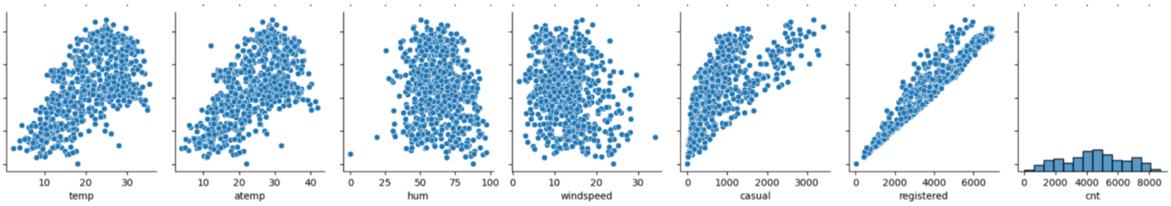
Ans. On looking at the pair-plot, registered has the highest correlation with cnt.



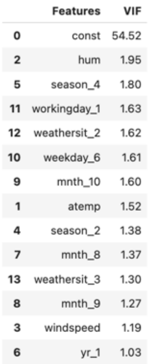
Q4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans4. The following are the assumptions of Linear Regression :

a) **The relationship between the independent and dependent variables should be linear :** Verified this using pair-plot and ensured that relationship between numerical variables with cnt is linear.

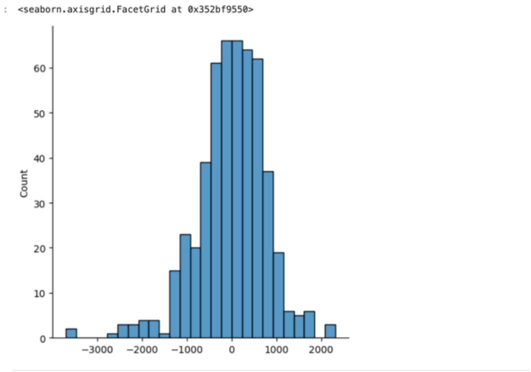


b) **Multicollinearity:** Validated that there is no multicollinearity by calculating VIF before predicting and ensuring that all VIF lie under value of 5. This is important because if values are collinear among themselves, it would increase redundancy in data and hence false results.



	Features	VIF
0	const	54.52
2	hum	1.95
5	season_4	1.80
11	workingday_1	1.63
12	weathersit_2	1.62
10	weekday_6	1.61
9	mnth_10	1.60
1	atemp	1.52
4	season_2	1.38
7	mnth_8	1.37
13	weathersit_3	1.30
8	mnth_9	1.27
3	windspeed	1.19
6	yr_1	1.03

c) **Homoscedasticity and normality of errors:** Verified by plotting a histogram of residual error. Resulted in normal distributed curve (mean around 0) and q-q plot to check the normality.



```
:  <seaborn.axisgrid.FacetGrid at 0x352bf9550>
```

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

General Subjective Questions

Q1. Explain the linear regression algorithm in detail. (4 marks)

Ans1. Linear regression is a supervised learning algorithm that computes linear relationship between dependent and independent variables by fitting it into a linear equation. It is of 2 types based on number of independent variables:

i. Simple Linear Regression: when data been provided has only one independent feature.
ii. Multiple Linear Regression: when data been provided has multiple independent features.

The goal of algorithm is to provide best fitted values of coefficients for equation :

$$y=\beta 0+\beta 1X1+\beta 2X2+.........\beta nXn$$

Some assumptions of linear regression are:

i. **Linearity**: There exists linear relationship between dependent and independent variable
ii. **Homoscedasticity**: the variance of the errors is constant across all levels.
iii. **Multicollinearity**: There is no high correlation between the independent variables
iv. **Normality**: the residuals are normally distributed

Steps of Linear Regression :

i. Data Preparation: involves cleaning data of inconsistences, null values and outliers. It also involves creating dummy variables for categorical variables and scaling the numerical variables.

ii. Fit Model : Next step after data preparation is to fit the model
iii. Check assumptions : Checking above stated assumptions like linearity, multicollinearity etc.
iv. Predict Values: Get predictions and make changes for p-values and vif of various features.
v. Residual Analysis and analysis on test set

Q2. Explain the Anscombe's quartet in detail. (3 marks)

Ans2. Anscombe's dataset highlights the importance of visualizing the data set instead of only relying on statistics results. Anscombe's quartet consists of 4 datasets each containing 11 x-y pairs and all four sets having same statistical properties of means, variance, R-squared, correlations, and linear regression lines. However, when these datasets were plotted as scatter plots, all of them had different representations.

Q3. What is Pearson's R ? (3 marks)

Ans3. The Pearson's R is a way of measuring a linear correlation between two variables. It is a number between –1 and 1 that indicates relationship between two variables.

Pearson's r represents the values as follows :

i. Value between 0 and 1 : Value of r between 0 and 1 indicates positive relationship between variables i.e. if one variable increases, the other increases too. The correlation increases as r gets closer to 1.
ii. r = 0 : indicates no relationship between the variables
iii. value between -1 and 0 : depicts a negative collinearity i.e. if one variable increases, other tends to decrease. The negative correlation between 2 variables increase as r gets close to value of -1.

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

Ans4. Scaling is a step of data preprocessing that involves scaling all variables to lie in a particular range. It is an important step as data may contain variables of a wide variety of ranges. When a linear regression model runs, it tries to weigh affect of every feature on the dependent variables. This weigh affect is calculated using coefficients of each feature. However, if the variable values are large, they would be multiplied with smaller coefficient thus indicating false importance of a variable. Therefore, scaling is an important step of data preprocessing. It improves stability and performance of the model

i. Normalized Scaling: Also known as MinMaxScaling, scales all the values to lie between [0,1]. It calculates scaled values using this formula:

x_new = (x-x_min) / (x_max – x_min)

It uses minimum and maximum values of feature to scale them. However, they do not handle outliers.

ii. Standardized Scaling: Standard deviation scales the feature by using mean and gaussian distribution. It calculates scaled values using this formula:

x_new = (x - mean) / std

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans5. VIF or variance inflation factor is used to indicate the correlation between features. This tends to break our linear regression assumption of multicollinearity. VIF value infinite suggests that the variables are strongly correlated.

(3 marks)

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression?

Quantiles are points taken at regular intervals from the cumulative distribution function (CDF) of a random variable. For example, the median is the 0.5 quantile because half the data lies below it.

A **Q-Q plot**, or **Quantile-Quantile plot**, used to assess whether a dataset follows a particular theoretical distribution or not. It compares the quantiles of the sample data with the quantiles of a theoretical distribution to see if they match, commonly used to check normality.

Understanding Q-Q Plot:

- **Straight Line**: If the data comes from the same distribution as the theoretical distribution, the points will roughly follow a straight line (45-degree angle).
- **Deviation from Line**: If the points deviate from the line, this indicates that the data differs from the theoretical distribution. The nature of the deviation can tell you about the type of difference. For example:
  - **S-Shape**: Data might have heavier tails (kurtosis) or lighter tails than the theoretical distribution.
  - **Upward or Downward Curving**: Indicates skewness; the data might be skewed left or right compared to the theoretical distribution.