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)**

Response:
Below are the inferences -

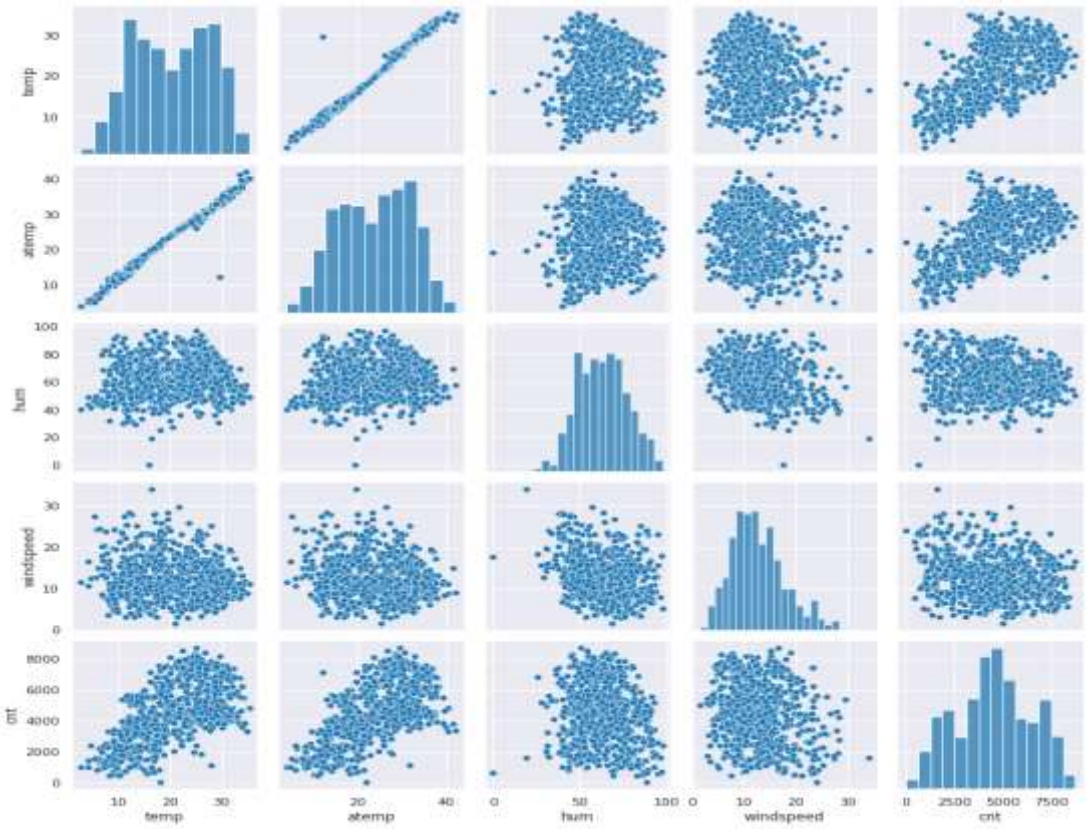
a. **Year:** There was significant jump in **Demand(cnt)** during 2019 compared to 2018
b. **Months: Demand(cnt)**is high during June, Aug & Sept months
c. **Season: Demand(cnt)**is high during Fall and Summer seasons and less during Spring season
d. **Workday & Holiday:** More **Demand(cnt)** is observed during Working days and Non holidays
e. **Weather Situation: Demand(cnt)** is high when weather is Clear, less or partly cloudy

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

Response:
To avoid redundancy dropping first column is useful. The importance or value of that left over variable can be found by remaining variables. It also help in avoid Dummy Variable trap.

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

Response:
If we consider all variables, then **registered** is highly correlated to target variable cnt (0.99). However, we have to drop it during data preparation to avoid multicollinearity. Both temp and atemp are correlated with target variable **cnt** (0.63)

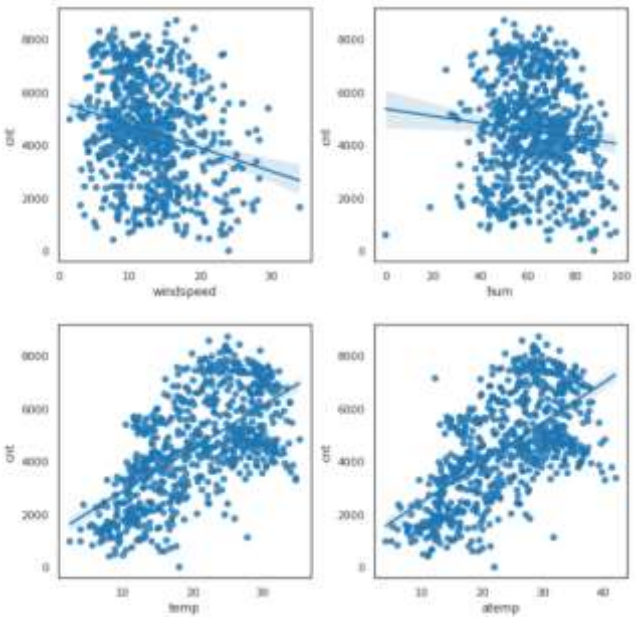


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Response:

Linear relationship

Linear relationship must exist between the dependant variable and the independent variables. Scatter/Regplot charts can help in visualizing and testing it.



No Multicollinearity

Multicollinearity's presence can adversely affect regression results. Independent variables shouldn't be highly correlated. This assumption can be validated with help of using VIF values. The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

$$VIF = \frac{1}{1 - R^2}$$

For interpreting the variance inflation factor:

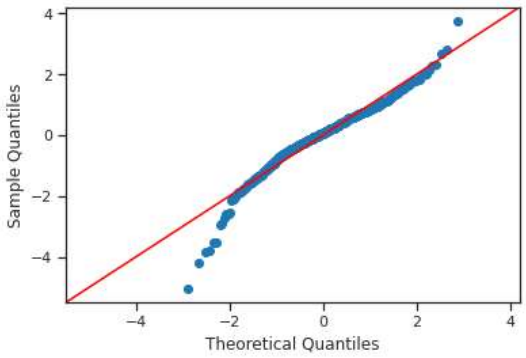
- 1 = very less or not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.

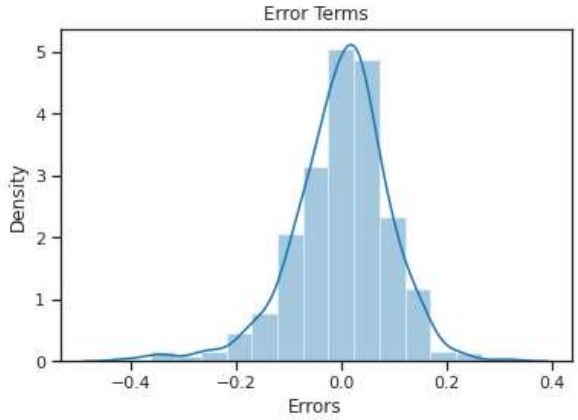
VIF for our final model

	Features	VIF
0	temp	4.768000
9	Workingday_Yes	4.049200
1	windspeed	3.440500
8	Year_2019	2.092700
7	Sat	1.668100
2	Summer	1.571200
5	Mist_Cloudy	1.531600
3	Winter	1.400100
6	Sept	1.200800
4	Light_Snow	1.083000

Multivariate Normal

All the variables should be multivariate normal or normally distributed. Also Normal distribution of error terms. This assumption can be checked using histogram of residuals v/s error terms or q-q plot Plots from the final model.





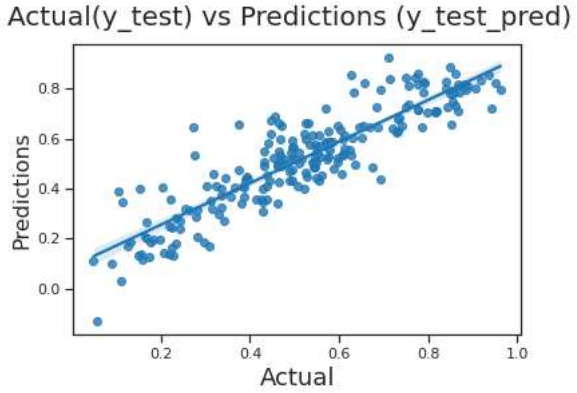
No Autocorrelation in the Data

There should be no autocorrelation in the data. The existence of autocorrelation can be verified with the Durbin-Watson test where it should score between 0 – 4 but less than 4, or scatter plot of residuals v/s time.

Durbin-Watson Test score of final model was ~2 (2.081) which is less than 4

Homoscedasticity

There should be homoscedasticity among the data. The scatterplot or displot can help in determining the homoscedasticity. Plot of final model



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

Response:
Based on the final model, below features are explaining the demand

a. Temperature (0.55)
b. weathersit (-0.28)
   Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
c. year (0.23) : 2019

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**
   Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range rather than trying to classify them into categories.
   There are two main types:
   a. Simple regression
      Simple linear regression uses traditional slope-intercept form, where m and b are the variables our algorithm will try to "learn" to produce the most accurate predictions. x represents our input data and y represents our prediction.

$$y = \alpha + \beta x,$$

      Or

      y=mx+b

   b. Multivariable regression
      A more complex, multi-variable linear equation might look like this, where w represents the coefficients, or weights, our model will try to learn.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \epsilon_i$$

      Or

      f(x,y,z)=w1x+w2y+w3z

2. **Explain the Anscombe's quartet in detail. (3 marks)**
   Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets. This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets

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

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables. The table below gives general rules of thumb:

Pearson correlation coefficient (*r*) value	**Strength**	**Direction**
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

The Pearson correlation coefficient (r) is one of several correlation coefficients that you need to choose between when you want to measure a correlation.
The Pearson correlation coefficient is a good choice when all of the following are true:

- Both variables are quantitative: You will need to use a different method if either of the variables is qualitative.
- The variables are normally distributed: You can create a histogram of each variable to verify whether the distributions are approximately normal. It's not a problem if the variables are a little non-normal.
- The data have no outliers: Outliers are observations that don't follow the same patterns as the rest of the data. A scatterplot is one way to check for outliers—look for points that are far away from the others.
- The relationship is linear: "Linear" means that the relationship between the two variables can be described reasonably well by a straight line. You can use a scatterplot to check whether the relationship between two variables is linear.

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

It is a step to pre-process data which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm**.**

Machine learning algorithms like linear regression, logistic regression, neural network, etc. that use gradient descent as an optimization technique require data to be scaled. The presence of feature value X in the formula will affect the step size of the gradient descent. The difference in ranges of features will cause different step sizes for each feature. To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model. Having features on a similar scale can help the gradient descent converge more quickly towards the minima.

Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

Normalisation	Standardisation
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

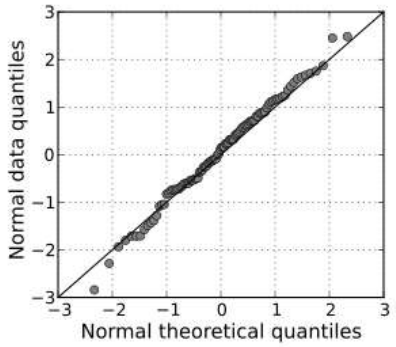
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Quantile-Quantile (Q-Q) plot is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution
This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.



Interpretations

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.
c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.
d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis