Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 1 goes below this line> (Do not edit)

1. `yr2019`: *Positively* has ***High*** demand
2. `workingday`: *Positively* has *Slight* demand
3. `season_winter`: *Positively* has *Slight* demand
4. `season_spring`: *Negatively* has *Slight* demand
5. `mnth_Jul`: *Negatively* has *Slight* demand
6. `weathersit_Mist`: *Negatively* has *Slight* demand
7. `weathersit_Snow`: *Negatively* has ***High*** demand

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)

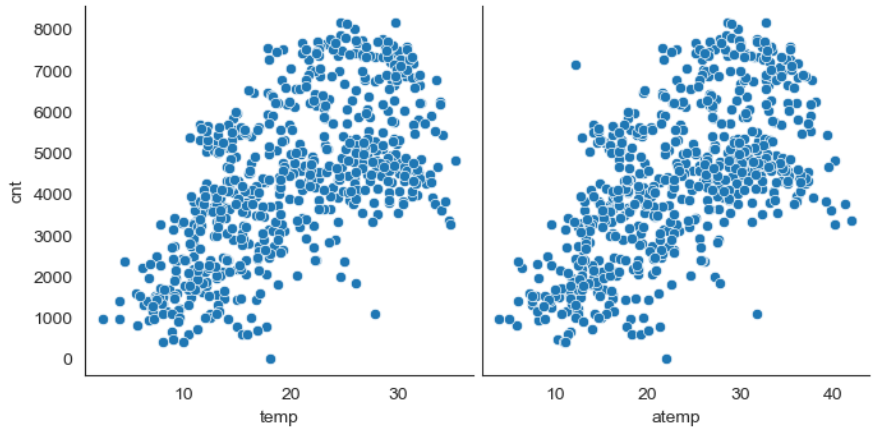
➔ *bk = pd.get_dummies(bk, dtype=int,* ***drop_first=True****)*

drop_first=True **->** This option will drop the first dummy variable created, in total (k-1) dummy variables will be created.

Dropping one of dummy variable will reduce the Multicollinearity (<VIF) between these dummies, as this information will be present in all other dummy variables together.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

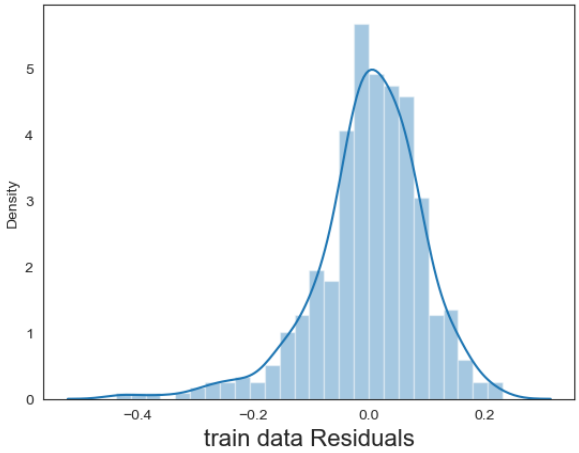
➔ The '`temp` & '`atemp`' has the highest correlation with the target variable.

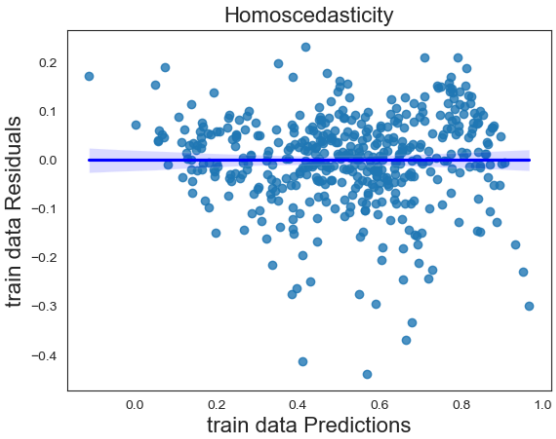


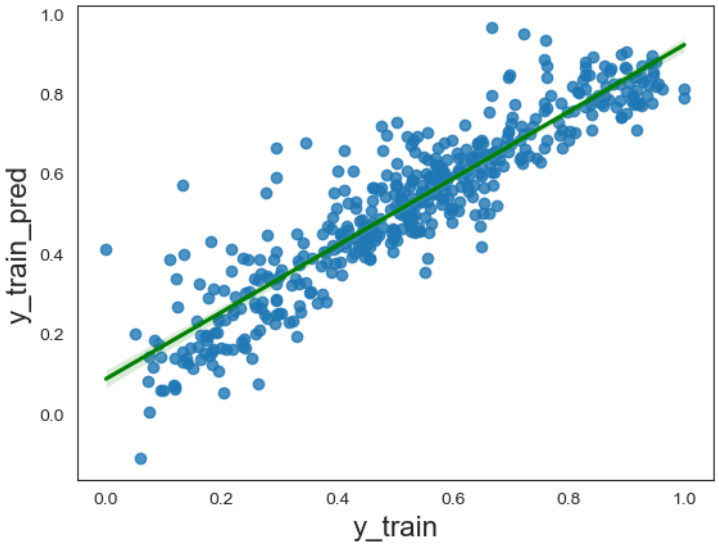
Pair-Plot

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

1. Train data Residual terms are following **Normal distribution**
2. **Homoskedasticity**, the variance of the train data Residual terms is almost Similar







3. The **VIF** value was minimized by removing the independent variables by RFE method
4. Remaining Independent variables removed by **P-values** analysis

➔ This seems to be a good Model that can very well *Generalize* various datasets

---

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

1. `temp:` *Positively* has ***High*** demand of bikes
2. `yr2019:` *Positively* has ***High*** demand of bikes
3. `weathersit_Snow:` *Negatively* has ***High*** demand of bikes
4. `hum:` *Negatively* has *Slight* demand of bikes
5. `windspeed:` *Negatively* has *Slight* demand of bikes

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

Linear Regression is Supervised Machine learning, finds the linearity relation between Explanatory/Independent variables and Predicted/Dependent variable.

Equation is derived using least error squares method, Residuals assumed to be followed Normal distribution for accurate Model.

1. ***Simple Linear Regression***: Machine learning using only Single independent variable

$$Y = m X + c$$

Y -> Dependent variable
m -> Slope
X -> Independent variable
c -> Intercept

2. ***Multiple Linear Regression***: Machine learning using only Multiple independent variable

$$Y=\beta 0+\beta 1X1+\beta 2X2+...+\beta pXp+\epsilon$$

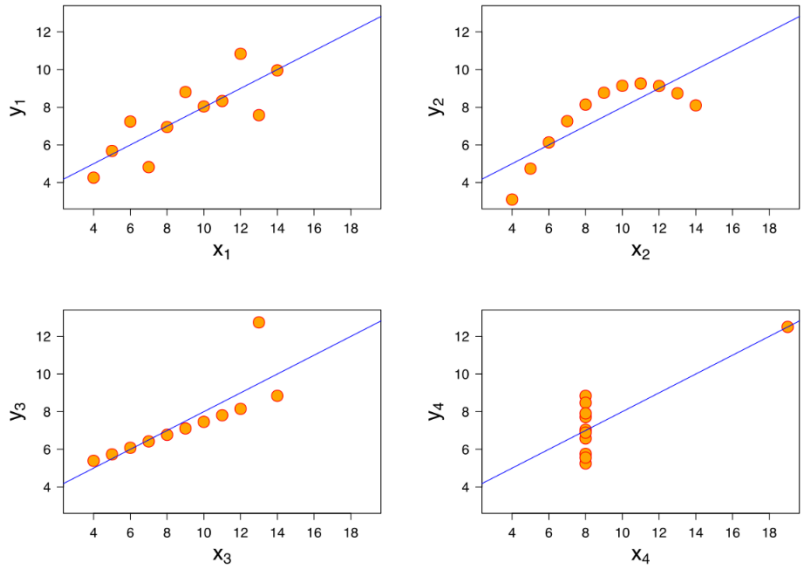
Y -> Dependent variable
X1 -> 1st Feature/Independent variable
Xp -> pth Feature/independent variable
β0 -> Intercept
β1 -> Co-efficient of X1
βp -> Co-efficient of Xp
ϵ -> Error

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

Anscombe's quartet consists of 4 identical data sets with properties like mean, variance, R-squared , correlation & linear regression but have different distributions.

Each set consists of 11 x-y pairs of data when plotted each distribution presents unique relation between x & y.



Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>

Pearson's R is a correlation coefficient measures the linear relation between two data sets.

The value of the coefficients is between 1 &-1, ratio between the covariance of the variables & product of their Standard deviations.

R = 1 : Perfectly linearity with positive slope
R = -1 : Perfectly linearity with Negative slope
R = 0 : No linear relation

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

Scaling is a technique performed as pre-proceesing of data before modeling, it standardise the data between the fixed range, the data may have different range of magnitude & units which may lead to incorrect modelling.

➔ The outliers introduces huge difference in this distribution, so the outlier removal is perfomed before Scaling.

1. ***Normalization/Min-Max Scaling***: distributes the data over range 0 to 1

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

2. ***Standardization Scaling***: distributes the data over their z-scores

$$x' = \frac{x - x_{mean}}{\sigma} \qquad z = \frac{x - \mu}{\sigma}$$

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

The value of **VIF** (Variance Inflation Factor) is infinite means there exists perfect multicollinearity/correlation between the independent variables, one or more variables has the same variance compared to others.

These variables can be dropped/eliminated for which the model linear regression to be optimized.

VIF = 1/(1-R^2) -> if the Independent variable completely described by the other variables then its correlation becomes '1' => *VIF = 1/(1-1) = 1/0 = Infinite*

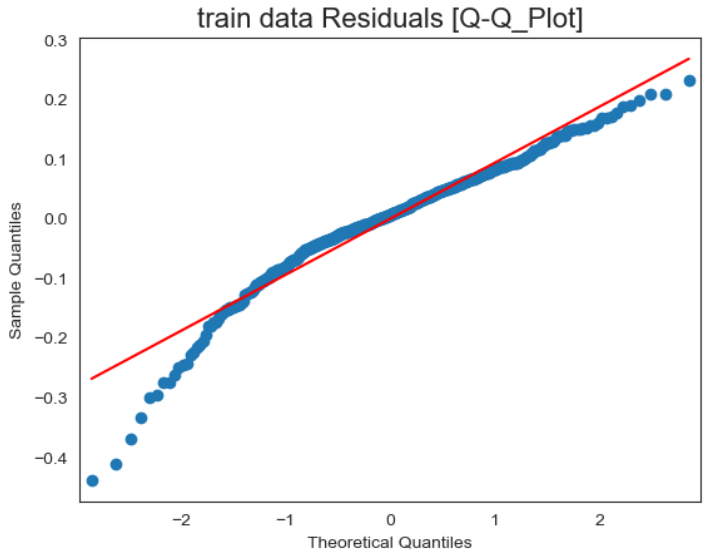
Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>

The **Quartile-Quartile** (Q-Q) Plots the Quartiles of Sample distribution with theoretical distribution to determine the data set follows any distribution such as Normal, Uniform or Exponential.

The Quartiles derived from same ditribution, the points should form a straight line.

The linear regression Q-Q Residuals plot points close to straight line indicates the Narmality, helps detrmining the Outliers & Skewness.



-> theoretical quantiles of a **Normal distribution**, are getting almost a *Straight* line..
