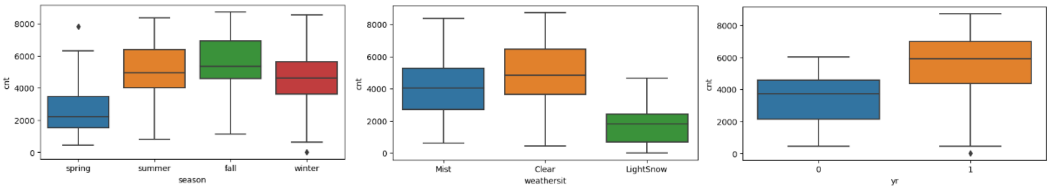
Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**
   season, weather situation & year seems to have strong effect on the dependent variable

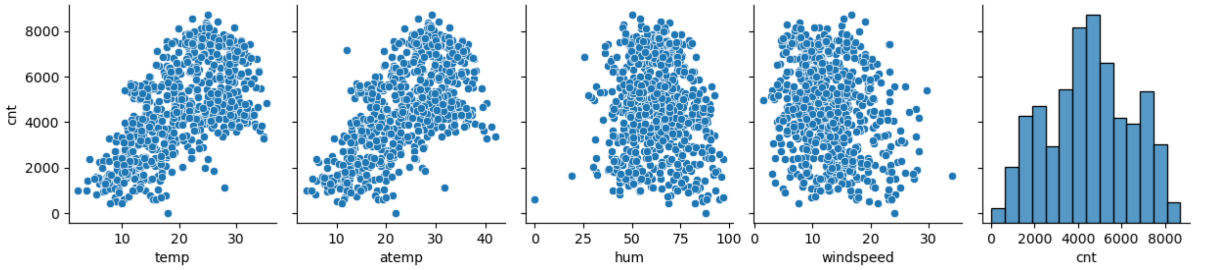


a. Dependent variable “cnt” is high when season is summer or fall and lowest in spring.
b. Dependent variable “cnt” is highest when weather clear and lowest when weather has light snow. Also, there is no record for heavy rain, which is understandable that no one would want to ride a bike in heavy rain.
c. Dependent variable “cnt” is higher in year 1 (2019) than previous year 0 (2018)

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   It is important to use drop_first=true during dummy variable creation, because we need to have only **n-1 dummy variables** out of total n distinct values (levels) of a categorical variable. When all other dummy variables have zero value it represents that the column we dropped had the value 1, so we don’t really need to have that column in out dataset, it would be redundant.

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

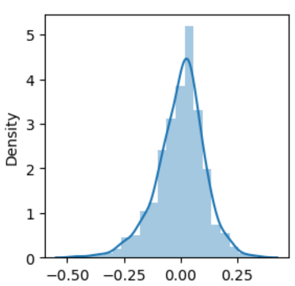
   temp and atemp seems to have higher correlation with cnt

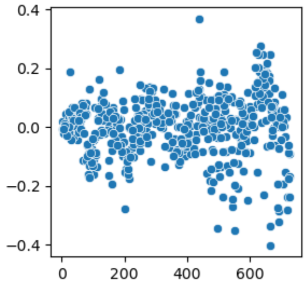


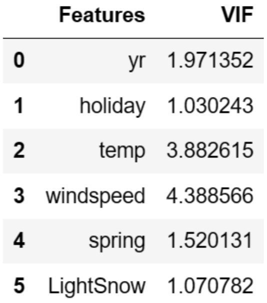
Since temp and atemp is basically the same attribute, we should consider only one. As per the data dictionary, **atemp** is the feeling temperature. **Feeling temperatur**e should be a better predictor of whether people would want to ride a bike or not, than the actual temperature.

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

   After building the model, we draw a distribution plot of the difference between actual values of Y in training set and the predicted values of Y from the training set (i.e. error terms). We also draw a scatter plot of these error terms to see if there is any clear pattern and if the variance at each value of x is same.







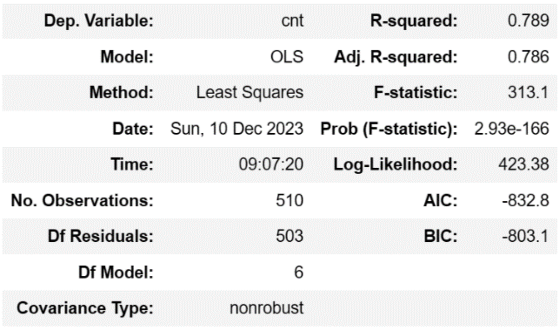
	Features	VIF
0	yr	1.971352
1	holiday	1.030243
2	temp	3.882615
3	windspeed	4.388566
4	spring	1.520131
5	LightSnow	1.070782

From the above graphs we can conclude:

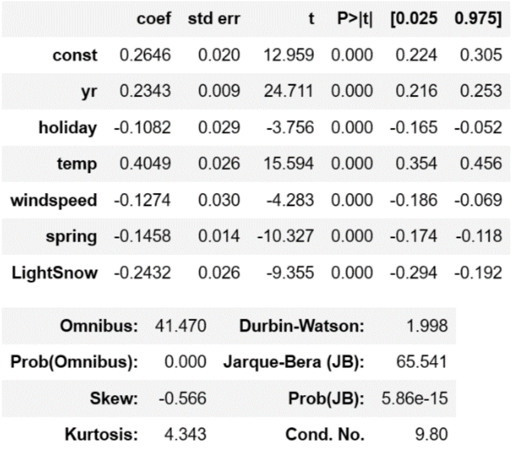
- Error terms are **normally distributed** around zero.
- There are no clear patterns, so we can say that Error terms are **independent** of each other. All VIFs are below 2. This means **Multicollinearity** is very less.
- The **Variance** at higher values of x is slightly more than at low values, but the difference is not too much. We can say its **Homoscedastic.**

We can say that in our model these assumptions of a linear regressions are holding true.

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



Dep. Variable:	cnt	**R-squared:**	0.789
Model:	OLS	**Adj. R-squared:**	0.786
Method:	Least Squares	**F-statistic:**	313.1
Date:	Sun, 10 Dec 2023	**Prob (F-statistic):**	2.93e-166
Time:	09:07:20	**Log-Likelihood:**	423.38
No. Observations:	510	**AIC:**	-832.8
Df Residuals:	503	**BIC:**	-803.1
Df Model:	6		
Covariance Type:	nonrobust		



	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2646	0.020	12.959	0.000	0.224	0.305
yr	0.2343	0.009	24.711	0.000	0.216	0.253
holiday	-0.1082	0.029	-3.756	0.000	-0.165	-0.052
temp	0.4049	0.026	15.594	0.000	0.354	0.456
windspeed	-0.1274	0.030	-4.283	0.000	-0.186	-0.069
spring	-0.1458	0.014	-10.327	0.000	-0.174	-0.118
LightSnow	-0.2432	0.026	-9.355	0.000	-0.294	-0.192

Omnibus:	41.470	**Durbin-Watson:**	1.998
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	65.541
Skew:	-0.566	**Prob(JB):**	5.86e-15
Kurtosis:	4.343	**Cond. No.**	9.80

Following are the top 3 features contributing significantly:

1. **temp** (Higher **Temperature** Increases the demand)
2. **weathersit** (When weather has **LightSnow** it Decreases the demand)
3. **yr** (Higher **Year** Increases the target, i.e. Year 2019 indicate higher demand for bike renting, as compared to 2018)

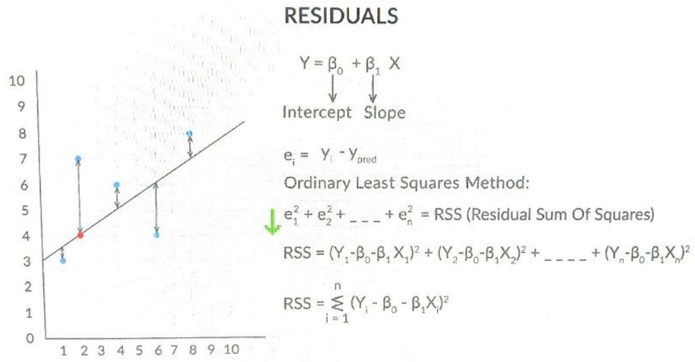
Using the coefficents we got from the model summary, we can make the equation of the best fitted line as

- ***cnt = 0.405(temp) - 0.243(LightSnow) + 0.234(yr) - 0.145(spring) - 0.127(windspeed) - 0.108(holiday) + 0.265***

General Subjective Questions

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

   Linear regression is a statistical method which creates a model for the relationship between a dependent variable (y) and one or more independent variables (x_1, x_2, …x_n). The model can explain the variation in the dependent variable, using the independent variables. Linear regression assumes that the relationship between the variables is linear, i.e. the change in the dependent variable is proportional to the change in the independent variables. Linear regression also assumes that the errors or residuals are normally distributed and independent of each other.



RESIDUALS

$$Y = \underset{\text{Intercept}}{\beta_0} + \underset{\text{Slope}}{\beta_1} X$$

$e_i = Y_i - Y_{pred}$

Ordinary Least Squares Method:

$e_1^2 + e_2^2 + \_\_\_ + e_n^2$ = RSS (Residual Sum Of Squares)

$$RSS = (Y_1 - \beta_0 - \beta_1 X_1)^2 + (Y_2 - \beta_0 - \beta_1 X_2)^2 + \_\_\_\_ + (Y_n - \beta_0 - \beta_1 X_n)^2$$

$$RSS = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

   1. Using a sample data we can create a regression line.
   2. Then calculate the residuals, i.e. difference between actual values of Y and predicted values of Y.
   3. Then we can sum these differences and square them to get the Residual Sum of Squares (RSS)
   4. This RSS is a representation of how best fit our regression line was. Even better representation is by calculating $\mathbf{R^2}$ as 1 – (RSS / TSS). Where TSS is Total Sum of Squares given by $(Y_1 – \bar{Y})^2 + …. + (Y_n – \bar{Y})^2$
   5. With this we can make multiple models and change them accordingly to minimise the error terms.
   6. Minimum the error in the regression line, the better our model can explain the dependent variable.

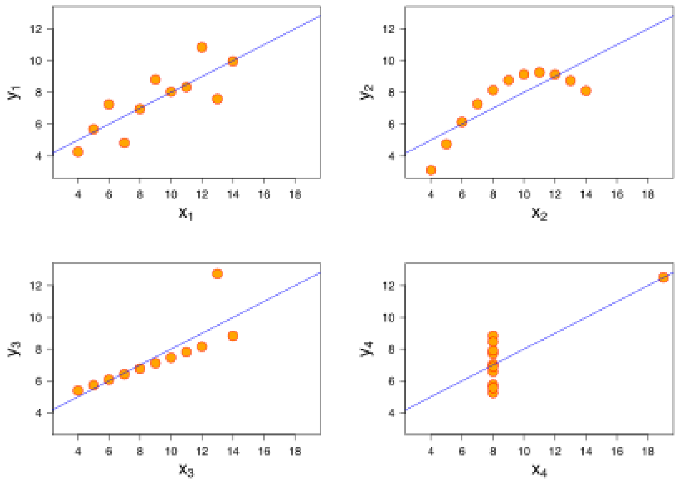
   Linear regression can be used for various purposes, such as finding the best-fit line for a set of data points, estimating the slope and intercept of the relationship, testing hypotheses about the coefficients, and making predictions based on new values of the independent variables.

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

Anscombe's quartet is a group of 4 datasets of x & y that have the same mean, standard deviation, and regression line, but they are qualitatively different. Since their distribution is different, they appear very different when plotted on a graph.

They demonstrate:

- The importance of graphing data when analysing it.
- The effect of outliers and other influential observations on statistical properties.



- The relationship between x and y in the first graph (top left) is linear and positive, meaning that y increases as x increases. The data points are close to the line of best fit, which suggests a strong correlation between the two variables.
- The second graph (top right) shows a nonlinear relationship between x and y, which cannot be captured by a simple linear regression. The data points form a curved pattern, indicating that y depends on x in a more complex way.
- The third graph (bottom left) has a linear relationship between x and y, but the line of best fit is distorted by an outlier. The outlier also reduces the correlation coefficient, which would otherwise be close to 1.
- The fourth graph (bottom right) illustrates how a single high-leverage point can create a misleading impression of a linear relationship between x and y. A high-leverage point can affect the slope of the line significantly. The correlation coefficient is high because of this point, but the rest of the data shows no clear pattern.

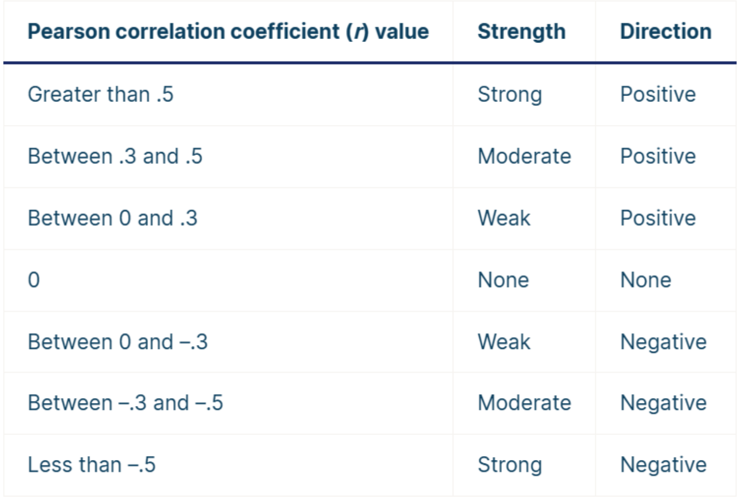
The quartet is often used to **illustrate the importance of looking at a set of data graphically** before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. What is Pearson’s R? (3 marks)

The **Pearson correlation coefficient (R)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

It summarizes the characteristics of a dataset (descriptive statistic). 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

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

Scaling is a process of transforming numerical data to a different range or scale.

Scaling is performed to make the data more comparable, reduce the effect of outliers, and improve the performance of a machine learning algorithms.

The two common types of scaling are:

- **Normalized scaling** rescales the data to a range between 0 and 1, where the minimum value becomes 0 and the maximum value becomes 1.
- **Standardized scaling** rescales the data to have a mean of 0 and a standard deviation of 1, where each value is subtracted by the mean and divided by the standard deviation.

Normalized scaling **preserves the shape** of the original distribution, while standardized scaling makes the distribution **more normal-like**.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**
   VIF indicates multicollinearity. The greater the VIF, the higher the degree of multicollinearity. So, in case when the regressor is **perfectly equal to a linear combination of other regressors**, the VIF tends to infinity.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**
   A Q-Q plot is a quantile-quantile plot. It is a graphical tool that compares the empirical distribution of the residuals (based on experiments and practical experience) of a linear regression model with the theoretical distribution of a normal random variable.
   A Q-Q plot plots the quantiles of the residuals against the quantiles of a standard normal distribution. If the residuals are normally distributed, then the points on the Q-Q plot should lie close to a straight diagonal line.
   This plot is useful for **checking the normality assumption** of linear regression, which is important for performing hypothesis tests and constructing confidence intervals for the model parameters.
   It can also help us **detect outliers, skewness, and heteroscedasticity** in the residuals.