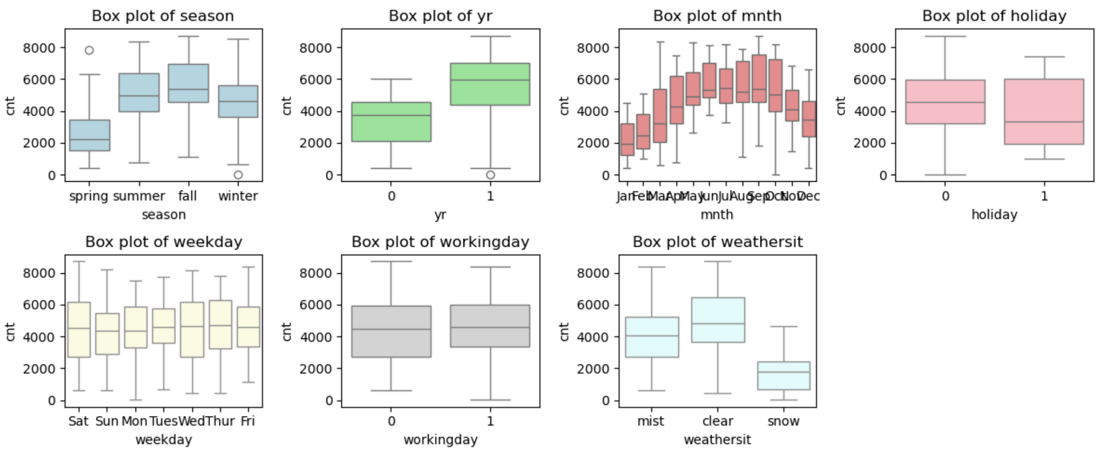
Assignment-based Subjective Questions

Senthil Kumar Selvam

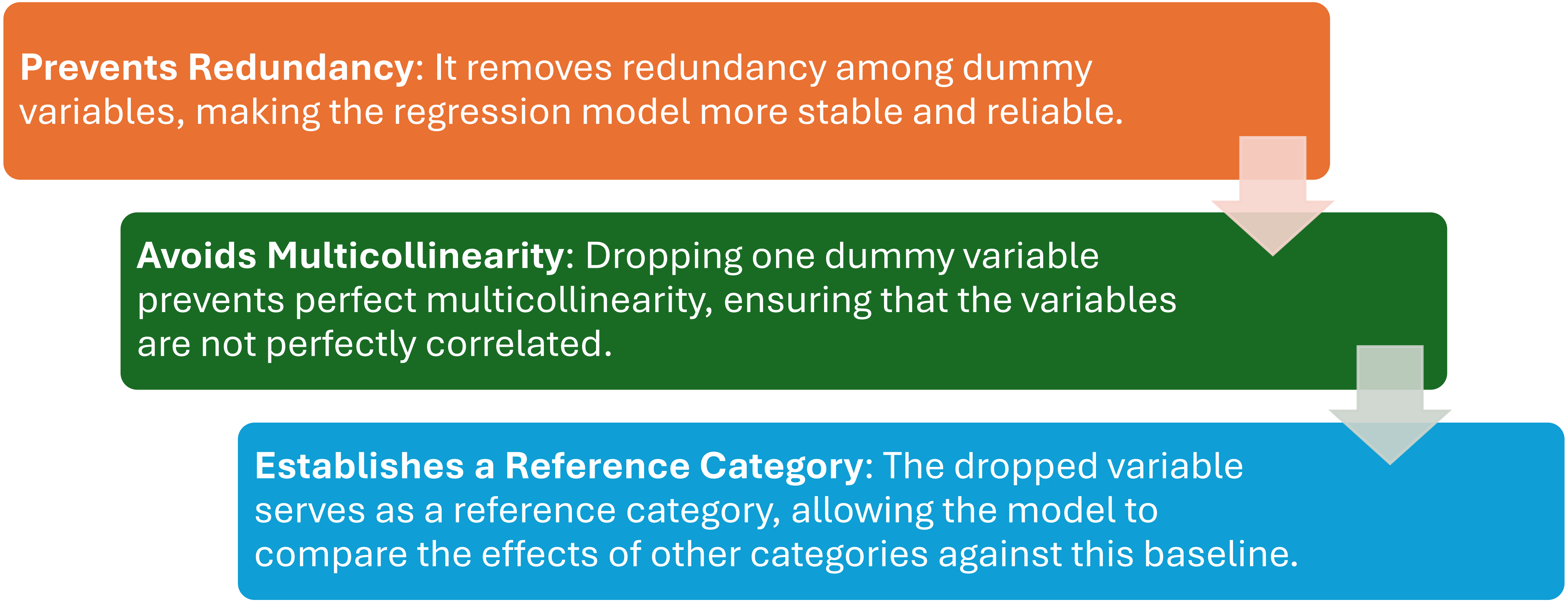
IIITB-Upgrad

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



Category	Observation
Season vs. Count (season vs. cnt)	The median count varies significantly across seasons. For example, Season 3 (fall) has the highest count.
Year vs. Count (yr vs. cnt)	The count is higher in the second year (yr = 1) compared to the first year. The median count is noticeably higher in the second year.
Month vs. Count (mnth vs. cnt)	There is a clear seasonal trend with counts increasing during middle months of the year (likely spring and summer) and decreasing in winter.
Holiday vs. Count (holiday vs. cnt)	The median count is slightly lower on holidays (holiday = 1) compared to non-holidays, but the overall distribution is similar, indicating that holidays don't drastically affect bike rentals.
Weekday vs. Count (weekday vs. cnt)	The median count remains relatively consistent across weekdays, suggesting that bike rentals do not vary much based on the day of the week.
Working Day vs. Count (workingday vs. cnt)	There is no significant difference between working days and non-working days. The distributions are similar, indicating bike rentals are consistent regardless of whether it's a working day or not.
Weather Situation vs. Count (weathersit vs. cnt)	The median count decreases as the weather situation worsens. Bad weather conditions result in the lowest count, showing that poor weather reduces bike rentals.

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



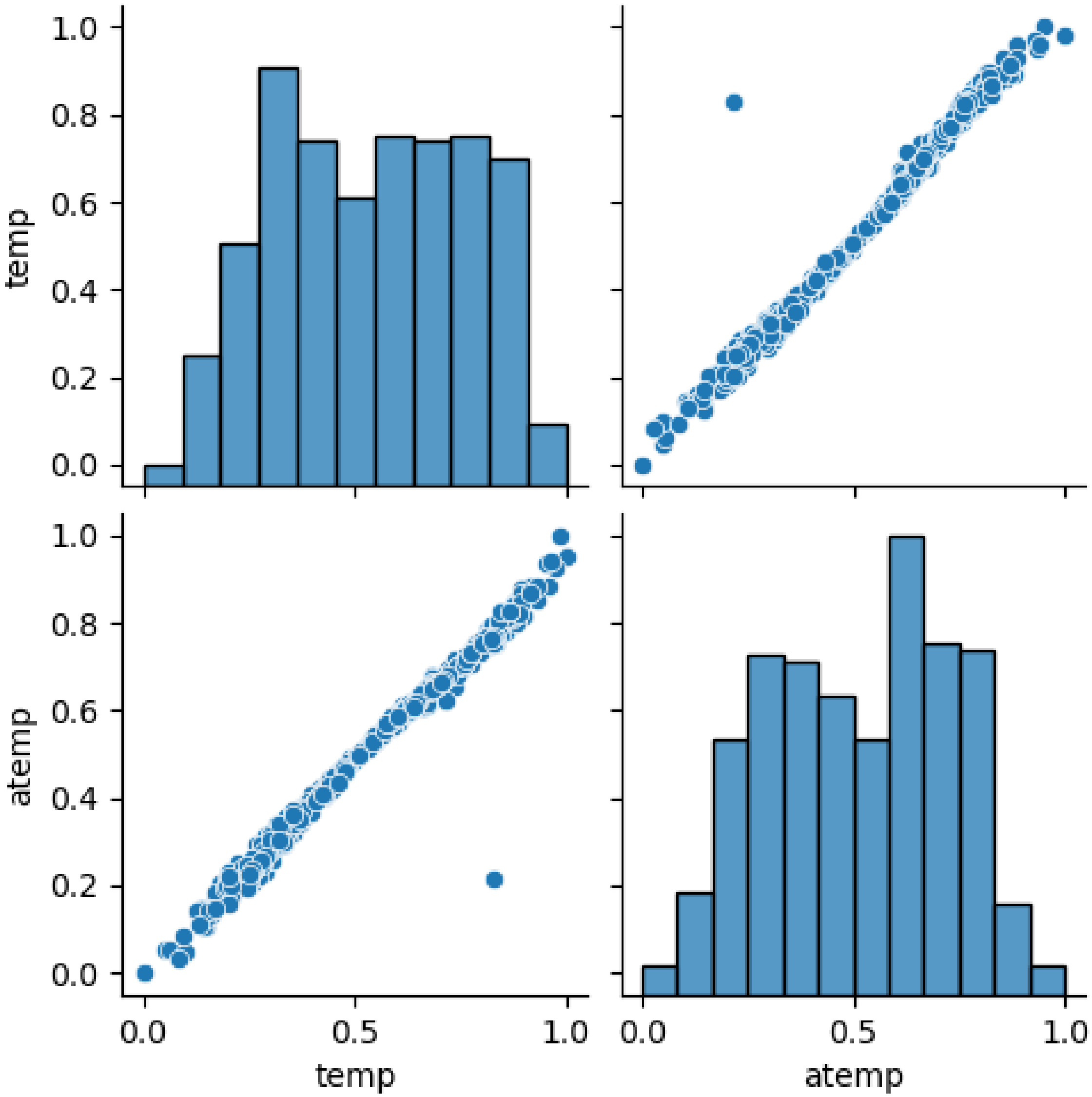
Prevents Redundancy: It removes redundancy among dummy variables, making the regression model more stable and reliable.

Avoids Multicollinearity: Dropping one dummy variable prevents perfect multicollinearity, ensuring that the variables are not perfectly correlated.

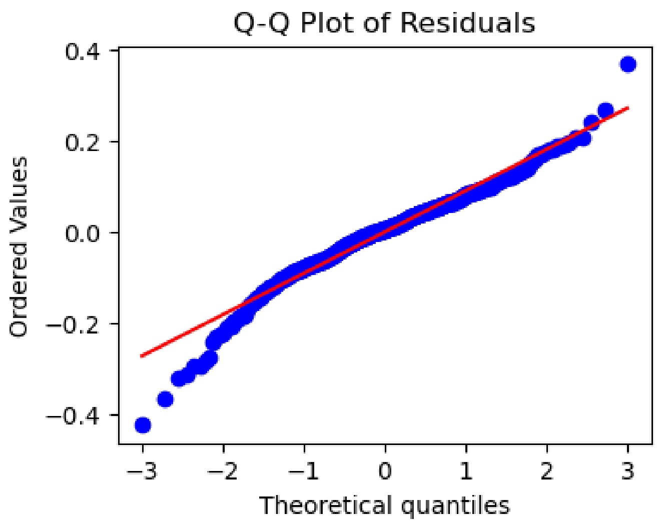
Establishes a Reference Category: The dropped variable serves as a reference category, allowing the model to compare the effects of other categories against this baseline.

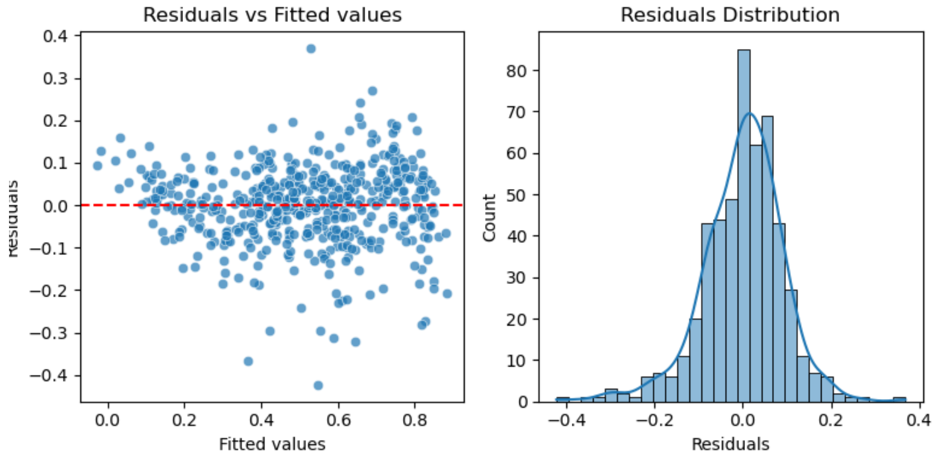
Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

- temp : temperature in Celsius vs atemp: feeling temperature in Celsius



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

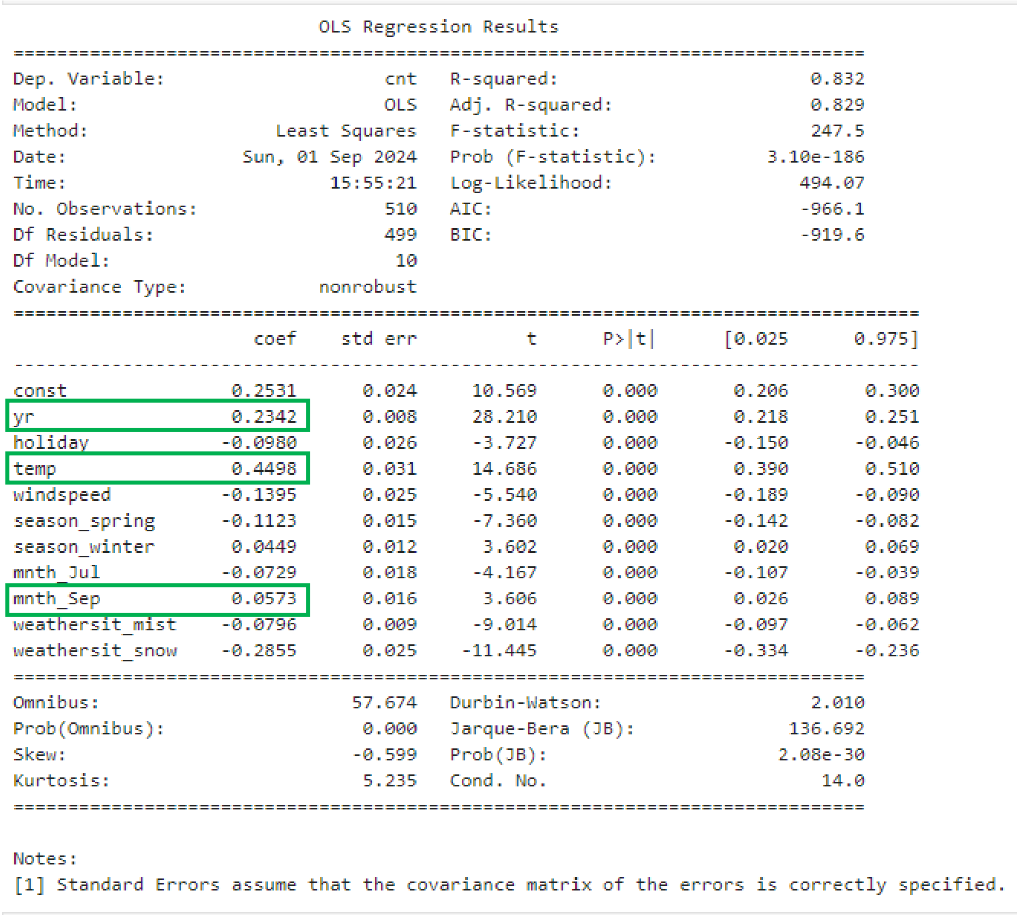




- **Normality Check**: Residuals are generally close to normal, as most points align with the red line.
- **Central Alignment**: Residuals near the center are well-distributed.
- **Model Fit:** R-squared, Adjusted R-squared, F-statistic, p < 0.0001)
- **Low VIF Values (VIF < 5):** Indicate low multicollinearity and are generally safe to include in the model.

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

Ranking	Variables	Correlation coefficient
1	temp	0.4498
2	yr	0.2342
3	Mnth_sep	0.0573



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.832
Model:                            OLS   Adj. R-squared:                  0.829
Method:                 Least Squares   F-statistic:                     247.5
Date:                Sun, 01 Sep 2024   Prob (F-statistic):          3.10e-186
Time:                        15:55:21   Log-Likelihood:                 494.07
No. Observations:                 510   AIC:                            -966.1
Df Residuals:                     499   BIC:                            -919.6
Df Model:                          10
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
const               0.2531      0.024     10.569      0.000       0.206       0.300
yr                  0.2342      0.008     28.210      0.000       0.218       0.251
holiday            -0.0980      0.026     -3.727      0.000      -0.150      -0.046
temp                0.4498      0.031     14.686      0.000       0.390       0.510
windspeed          -0.1395      0.025     -5.540      0.000      -0.189      -0.090
season_spring      -0.1123      0.015     -7.360      0.000      -0.142      -0.082
season_winter       0.0449      0.012      3.602      0.000       0.020       0.069
mnth_Jul           -0.0729      0.018     -4.167      0.000      -0.107      -0.039
mnth_Sep            0.0573      0.016      3.606      0.000       0.026       0.089
weathersit_mist    -0.0796      0.009     -9.014      0.000      -0.097      -0.062
weathersit_snow    -0.2855      0.025    -11.445      0.000      -0.334      -0.236
==============================================================================
Omnibus:                       57.674   Durbin-Watson:                   2.010
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              136.692
Skew:                          -0.599   Prob(JB):                     2.08e-30
Kurtosis:                       5.235   Cond. No.                         14.0
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

General Subjective Questions

Senthil Kumar Selvam

IIITB-Upgrad

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

Linear regression	computes the linear relationship between the dependent variable (X) and one or more independent features or variables(y) by fitting a linear equation to observed data.
Ways of LR	Simple Linear Regression : only one independent feature / Multiple Linear Regression: more than one feature Univariate Linear Regression : only one dependent variable/ Bi Variate Linear Regression : there are more than one dependent variables,
Data Collection and Processing	Dependent variable (Y) and Independent Variable (X) Handling Missing Data Data Transformation Splitting Data
Exploratory Data Analysis (EDA)	Visualize Relationships
Model Selection and Training	Simple Linear Regression and Multiple Linear Regression Fit the Model Cost Function (Error Calculation) Optimization (Gradient Descent)
Model Evaluation	Residual Analysis R-squared Value Adjusted R-squared P-Values and Significance Tests
Model Validation	Testing on Test Data: MSE, RMSE, or MAE. Cross-Validation
Model Interpretation and Deployment	Coefficient Interpretation Predictive Power Final Model Selection

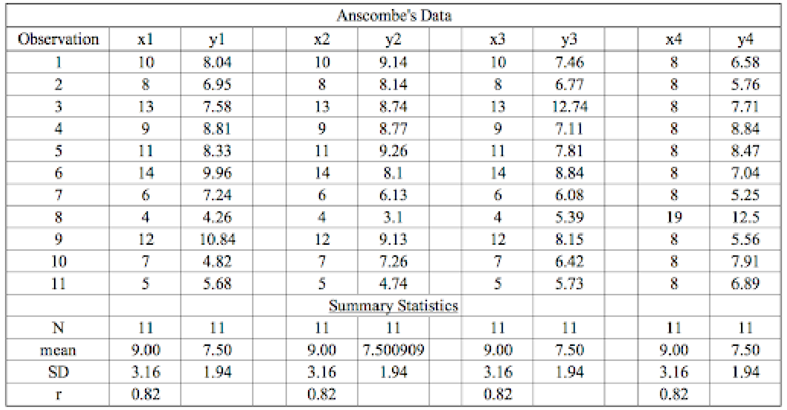
(Contd...,)Explain the linear regression algorithm in detail. (4 marks)

Aspect	Simple Linear Regression	Multiple Linear Regression	Expected Outcome
Model Equation	$y=\beta0+\beta1x+\epsilon$	$y=\beta0+\beta1x1+\beta2x2+\cdots+\beta nxn+\epsilon$	Predict a linear relationship between the dependent variable and predictors
Dependent Variable	y	y	Target variable to predict
Independent Variables	Single predictor x	Multiple predictors x1,x2,...xn	Variables used to predict y
Coefficients	β0: Intercept β1: Slope	β0: Intercept β1,β2,...,βn: Slopes	Measure the strength and direction of the relationship between predictors and y
Error Term	ϵ	ϵ	Represents the difference between observed and predicted values
Assumptions	- Linearity - Independence - Homoscedasticity - Normality	- Linearity - Independence - Homoscedasticity - Normality	Ensure valid model inferences and accurate predictions
Fitting Method	Least Squares	Least Squares	Minimize the sum of squared errors between observed and predicted values
Evaluation Metrics	- R-squared - Mean Squared Error (MSE) - Root Mean Squared Error (RMSE)	- R-squared - Mean Squared Error (MSE) - Root Mean Squared Error (RMSE	Assess model accuracy and fit
Applications	Predicting relationships between two variables	Predicting relationships between a dependent variable and multiple independent variables	Estimate and interpret the effects of predictors on the target variable

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

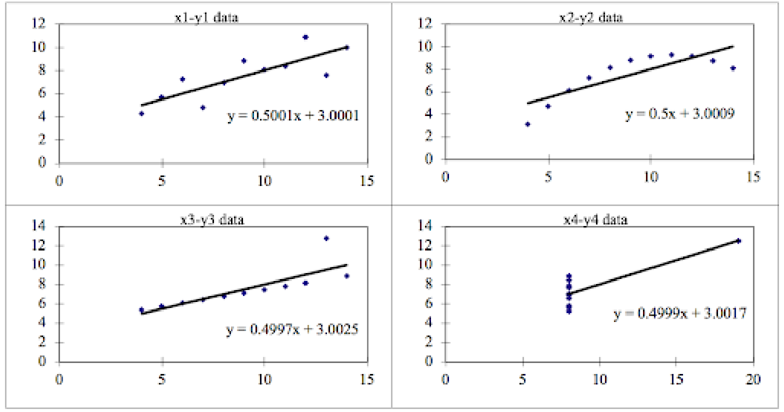
Ref: https://builtin.com/data-science/anscombes-quartet

- Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different greatly when graphed.
- It is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties.
- This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data.
- **Summary of Insights**
- **Dataset I**: A typical linear relationship suitable for linear regression.
- **Dataset II**: A parabolic (nonlinear) relationship not suited for linear regression.
- **Dataset III**: A linear relationship influenced by an outlier, highlighting the impact of outliers on analysis.
- **Dataset IV**: A vertical cluster with one outlier, leading to misleading statistical interpretations.



Anscombe's Data

Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	



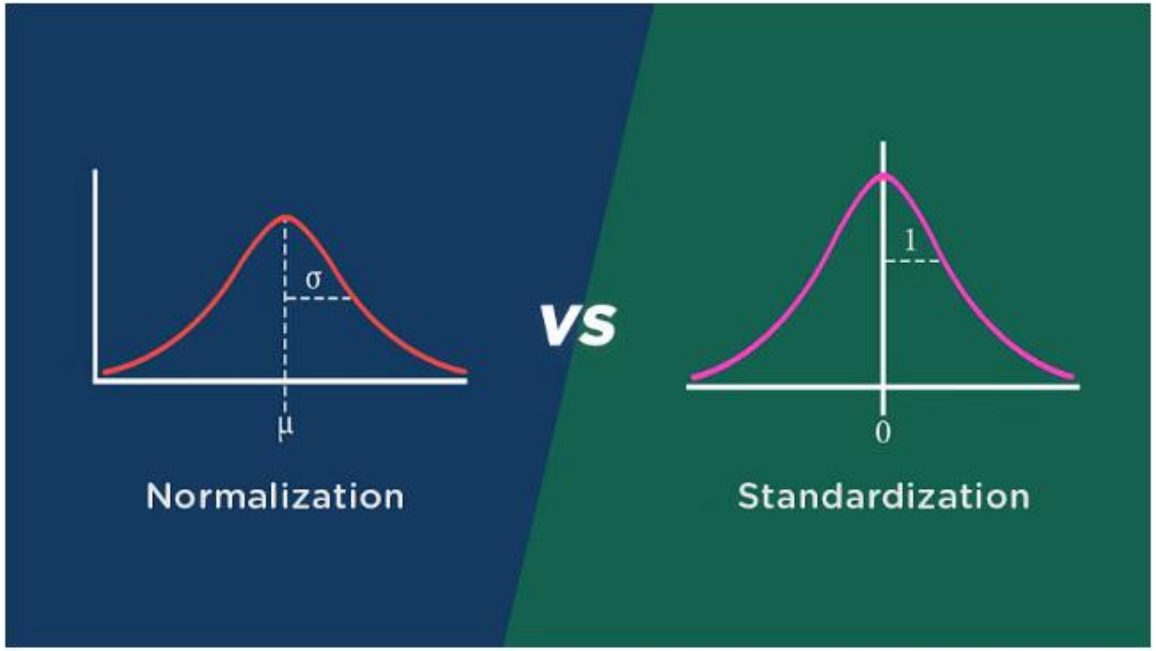
What is Pearson's R? (3 marks)

- 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.

Aspect	**Description**	**Expected Outcome**
Range	$-1 \leq r \leq 1$	r will fall between -1 and 1, indicating the strength and direction of the relationship.
Interpretation of r	- $r=1$: Perfect positive linear correlation - $r=-1$: Perfect negative linear correlation - $r=0$: No linear correlation	Determine if the relationship is perfectly linear (positive/negative) or non-existent.
Direction of Relationship	- $r>0$: Positive linear relationship - $r<0$: Negative linear relationship	Identify whether the variables increase together or one increases while the other decreases.
Key Characteristics	- Measures linear relationship - Sensitive to outliers - Symmetrical: $r(x,y)=r(y,x)$ - Unitless measure	Interpret r carefully, considering potential outliers, symmetry, and lack of units.
Applications	- Assessing relationships between variables in fields like psychology, economics - Exploratory data analysis	Use r to identify and quantify linear relationships during data analysis.

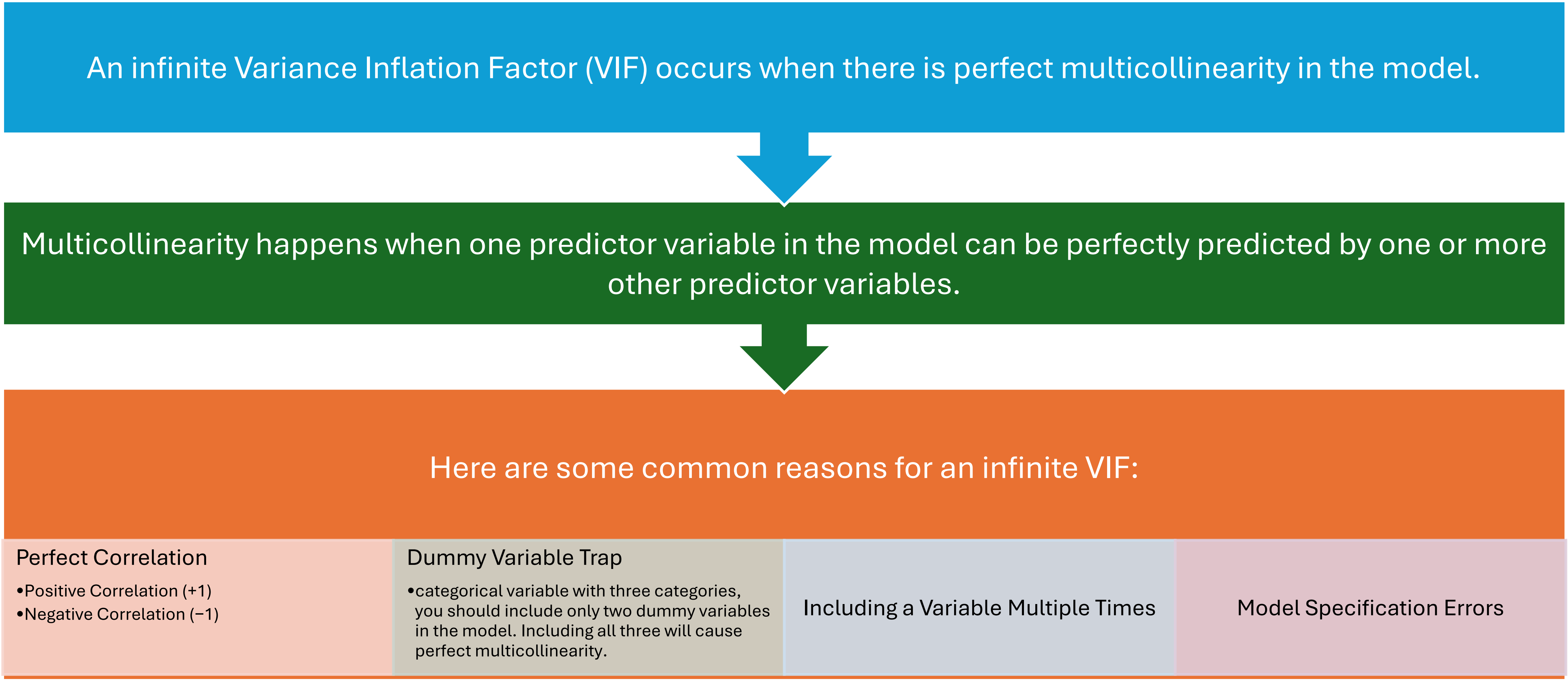
What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

- Scaling
  - *step of data Pre-Processing 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.*
- Reason for Scaling
  - Improving Model Performance
  - Accelerating Convergence
  - Ensuring Uniform Influence
  - Distance-Based Algorithms



Normalization	Standardization
This method scales the model using minimum and maximum values.	This method scales the model using the mean and standard deviation.
When features are on various scales, it is functional.	When a variable's mean and standard deviation are both set to 0, it is beneficial.
Values on the scale fall between [0, 1] and [-1, 1].	Values on a scale are not constrained to a particular range.
Additionally known as scaling normalization.	This process is called Z-score normalization.
When the feature distribution is unclear, it is helpful.	When the feature distribution is consistent, it is helpful.

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



An infinite Variance Inflation Factor (VIF) occurs when there is perfect multicollinearity in the model.

Multicollinearity happens when one predictor variable in the model can be perfectly predicted by one or more other predictor variables.

Here are some common reasons for an infinite VIF:

Perfect Correlation

- Positive Correlation (+1)
- Negative Correlation (−1)

Dummy Variable Trap

- categorical variable with three categories, you should include only two dummy variables in the model. Including all three will cause perfect multicollinearity.

Including a Variable Multiple Times

Model Specification Errors

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

- A Q-Q (Quantile-Quantile) plot
  - is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, most commonly the normal distribution.
  - It plots the quantiles of the dataset against the quantiles of the theoretical distribution.
  - If the data follows the theoretical distribution, the points on the Q-Q plot will approximately lie on a straight line.

Aspect	Use	Importance
Checking Normality of Residuals	Compares residuals to a normal distribution.	Validates a key assumption of linear regression.
Identifying Deviations	Identifies skewness, heavy tails, or other departures.	Helps diagnose issues affecting model accuracy.
Model Diagnostics and Improvement	Suggests adjustments like transformations or outlier handling	Enhances model reliability and predictive power.
Assumption Verification	Confirms normality assumption for residuals.	Ensures valid hypothesis testing and confidence intervals.
Outlier Detection	Highlights extreme deviations.	Identifies and addresses outliers influencing the model.

Thank you..!

Senthil Kumar Selvam

IIITB-Upgrad