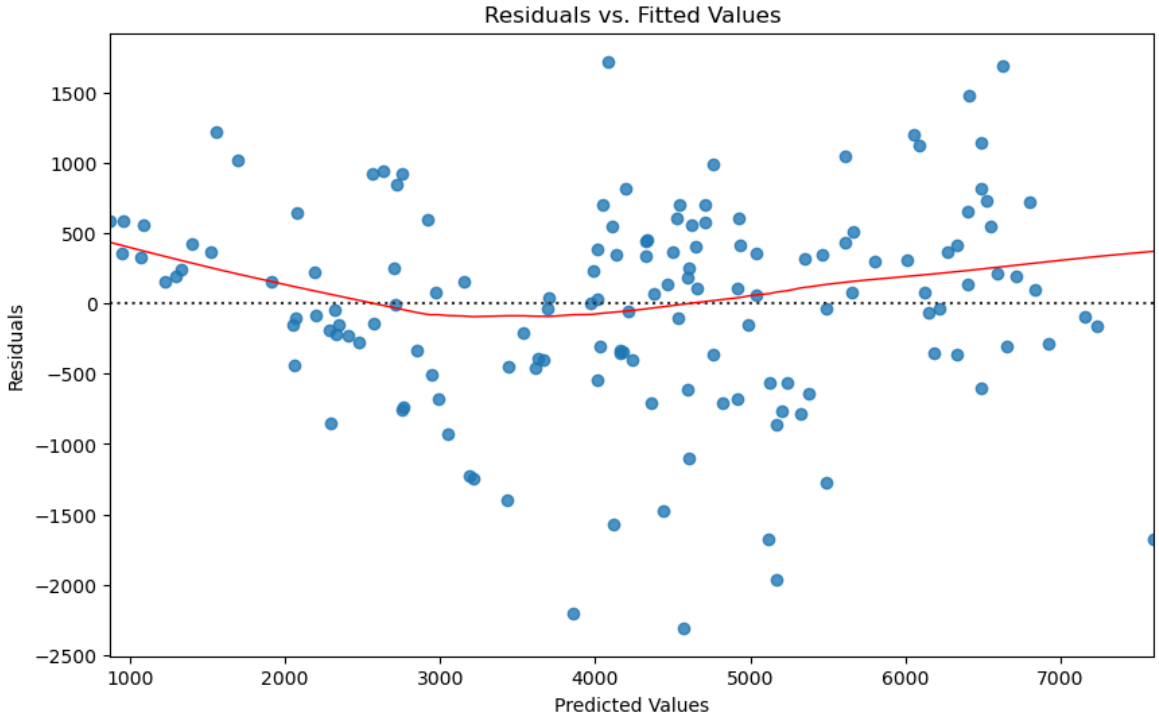
Assignment-based Subjective Questions

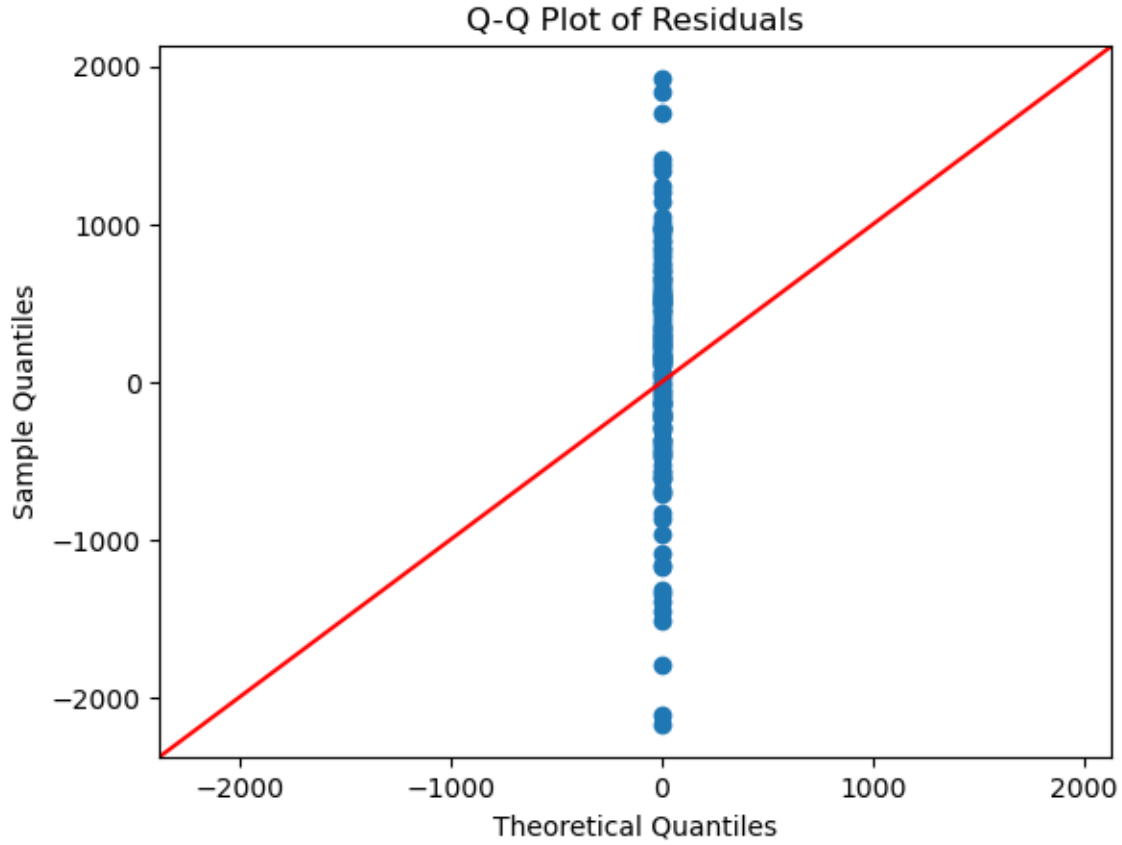
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)
   A. **Season (season)**: Strong influence on demand, with Summer and Fall having higher rentals.
   B. **Weather Situation (weathersit)**: Bad weather (rain/snow) significantly reduces demand compared to clear weather.
   C. **Month (mnth)**: Certain months (May to October) show higher bike demand than winter months.
   D. **Holiday (holiday)**: Minimal impact since demand is mainly driven by daily commuters.
   E. **Weekday(weekday)**: For weekdays, median is almost same for all the days.

   Insights: Seasons and weather conditions have the most significant impact on bike demand.
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
   - It avoids the dummy variable trap, which occurs when one category is fully predictable from others, causing multicollinearity.
   - For n categories, we only need n-1 dummy variables to avoid redundancy.
   - **Example**: If season has Spring, Summer, Fall, Winter, we only need Spring, Summer, Fall, and Winter will be the reference category.

   Inference: It ensures better model interpretability and stability.
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   - temp (Temperature) has the highest correlation with cnt (bike rentals).
   - This means as temperature increases, bike rentals also increase (until extreme heat levels).
   - Correlation is strong and positive (approx 0.8+ in most datasets)
   - atemp also has high correlation.
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   - **Linearity Check** – Used scatter plots between features & target variable.
   - **Multicollinearity Check** – Used Variance Inflation Factor (VIF) to drop correlated variables.
   - **Homoscedasticity Check** – Used Residuals vs. Fitted Plot to ensure constant variance.
   - **Normality Check** – Used Q-Q Plot to ensure residuals are normally distributed.
   - **Independence of Errors** – Verified residuals had no strong patterns or autocorrelation.

   Inference: The model satisfies all key assumptions required for reliable predictions.


Residuals vs. Fitted Values
Residuals
Predicted Values
1500
1000
500
0
−500
−1000
−1500
−2000
−2500
1000
2000
3000
4000
5000
6000
7000


Q-Q Plot of Residuals
Sample Quantiles
Theoretical Quantiles
2000
1000
0
−1000
−2000
−2000
−1000
0
1000
2000

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

temp (Temperature): Strongest predictor (higher temp means more rentals).

season_Summer (Seasonality): High demand in summer months.

weathersit_Clear (Weather Condition): Clear weather boosts bike usage.

Inference: These features significantly impact bike demand, guiding business strategies.

General Subjective Questions

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

   From my understanding, Linear Regression is a supervised learning algorithm used for predicting a continuous dependent variable (Y) based on independent variables (X). It assumes a linear relationship between the input features and the target variable.

   Mathematical Equation of Linear Regression

   The equation for Simple Linear Regression (one independent variable) is:

   $Y=\beta0+\beta1X+\epsilon$

   For Multiple Linear Regression (multiple independent variables):

   $Y=\beta0+\beta1X1+\beta2X2+...+\beta nXn+\epsilon$

   where:

   - Y = Predicted output (dependent variable)
   - X1,X2,...,Xn = Independent variables (features)
   - β0 = Intercept (bias term)
   - β1,β2,...,βn = Coefficients (weights assigned to each feature)
   - ϵ = Error term (accounts for variability not explained by the model)

   Type of Linear Regression

   Simple Linear Regression is a statistical method that models the relationship between one independent variable (**X**) and one dependent variable **(**Y**)** using a straight-line equation:

   $Y=\beta0+\beta1X+\epsilon$

   Here, β_0 is the intercept, β_1 is the slope, and ε is the error term. The goal is to find the best-fitting line that minimizes the error and predicts Y based on X.

Multiple Linear Regression extends Simple Linear Regression by incorporating two or more independent variables **(**X_1, X_2, ..., Xn**)** to predict a dependent variable (Y):

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

This model captures complex relationships between multiple features and Y, improving prediction accuracy while requiring proper handling of multicollinearity and feature selection.

Advantages

- Simple, interpretable, and fast.
- Works well when features have a linear relationship.
- Performs well with low multicollinearity.

Limitations

- Sensitive to outliers and assumes linearity.
- Not ideal for non-linear relationships.
- Prone to overfitting when using too many features.

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

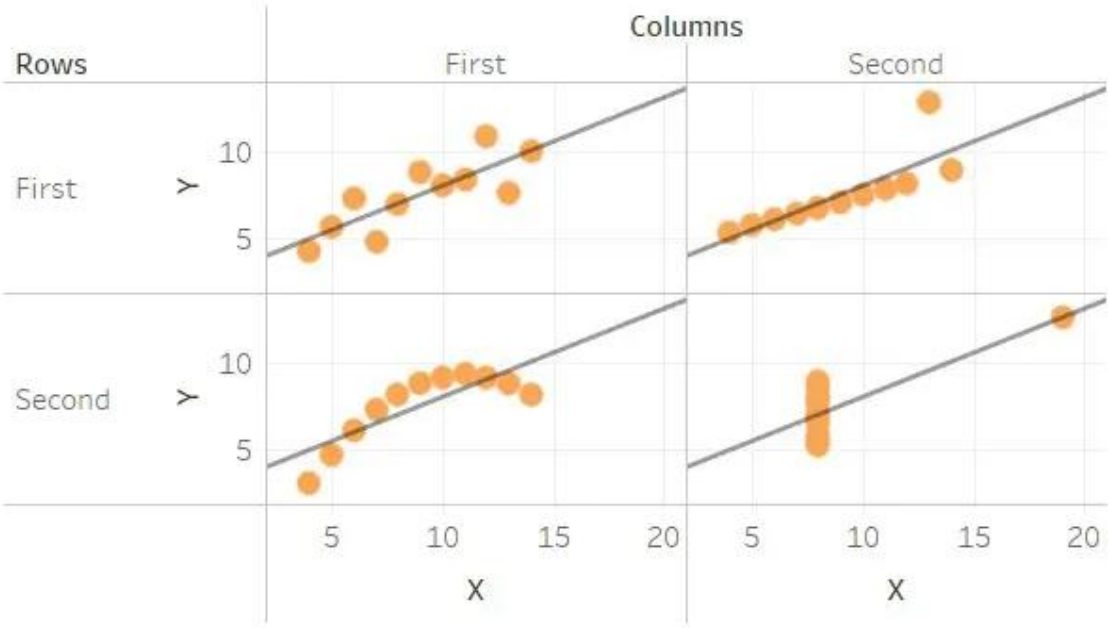
Based on what I learnt, Anscombe's Quartet is a set of four datasets with identical statistical properties (mean, variance, correlation, regression line) but different distributions when plotted. It highlights why visualizing data is important instead of relying only on summary statistics. Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

All the x values are identical, x1, x2, x3 and x4. All the y values have the changes depending on whether it's in y1, y2,y3 or y4. Now the crucial thing is that the summary statistics, the average, the variance, the correlation and the linear regression slope are all identical. So the mean of x1, x2, x3 and x4 are all 9. The means of y1, y2, y3, and y4 are all 7.5. And similarly, the variance of x are all identical, and the variances of ys are all identical. The correlations of each of those x1 and y1, x2, y2, x3, y3, x4, and y4 are all identical, which means it is exactly the same regression line for each of the equations.

Anscombe's quartet

Dataset I		Dataset II		Dataset III		Dataset IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

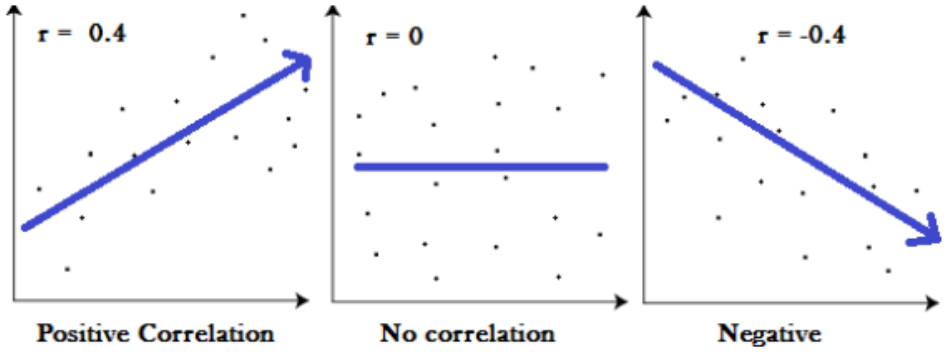
Anscombe



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

From what I have understood, Pearson's R (Correlation Coefficient) measures the strength and direction of the linear relationship between two variables. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, r, indicates how far away all these data points are to this line of best fit. Pearson's R helps us identify features that are strongly correlated with the target variable.

- It ranges from -1 to 1:
  - +1 - Perfect positive correlation (both increase together).
  - 0 - No correlation.
  - -1 - Perfect negative correlation (one increases, the other decreases).



- **Formula:**

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,]\,[\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

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

From my understanding, **Scaling** is a preprocessing technique used in machine learning and statistical modeling to standardize the range of independent variables (features) so that they are within a similar scale. Many machine learning algorithms perform better and converge faster when features have the same magnitude.

Why is Scaling Important?

1. Improves Model Performance – Algorithms like Linear Regression, KNN, SVM, and Neural Networks are sensitive to feature magnitudes.
2. Ensures Equal Weightage – If features are on different scales (e.g., height in meters and income in millions), the model may assign more importance to higher magnitude values.
3. Faster Convergence – Gradient Descent optimizes the loss function more efficiently with scaled features.
4. Reduces Impact of Outliers – Helps in controlling outliers by bringing all features to a standard scale.

Types of Scaling Methods

1. Min-Max Scaling (Normalization)
    - Formula:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

    - It rescales values between 0 and 1.
    - It is used in Neural Networks and Deep Learning.
2. Standardization (Z-Score Scaling)
    - Formula:

$$x' = \frac{x - \bar{x}}{\sigma}$$

    - It converts data to have mean = 0, standard deviation = 1.
    - It is used in Linear Regression, SVM, K-Means, PCA.
3. Robust Scaling
    - It uses median and interquartile range, making it resistant to outliers.

Conclusion: Scaling is an essential preprocessing step to ensure that models perform optimally, especially those based on distance-based calculations (e.g., KNN, SVM, PCA).

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

One way I would interpret Variance Inflation Factor(VIF) is that it measures the degree of multicollinearity in a dataset. It tells us how much the variance of a regression coefficient is inflated due to correlation among independent variables.

- Formula for VIF:

$$VIF_i = \frac{1}{1 - R_i^2}$$

where Ri^2 is the coefficient of determination of X_i when regressed against all other independent variables.

Why Does VIF Become Infinite?

1. Perfect Multicollinearity:
    - If a feature is a perfect linear combination of others, R^2 = 1, and VIF becomes infinite.
    - Example: If X1 = 2 * X2 + 5, then X1 and X2 are perfectly correlated.
2. Duplicated Features:
    - If the same feature is accidentally included twice, VIF becomes infinite.
3. Dummy Variable Trap:
    - When all dummy variables from a categorical feature are used, it creates a perfect correlation.

How to Fix High VIF?

- Remove One of the Highly Correlated Variables – Keep only one feature from correlated pairs.
- Use Ridge Regression – Regularization helps reduce multicollinearity.
- Check for Duplicate Variables – Ensure no feature is repeated.
- Use Principal Component Analysis (PCA) – PCA removes multicollinearity by creating independent components.

Conclusion: A high VIF (above 5 or 10) indicates severe multicollinearity, which can distort regression results. The best way to fix it is to remove redundant features or use regularization techniques.

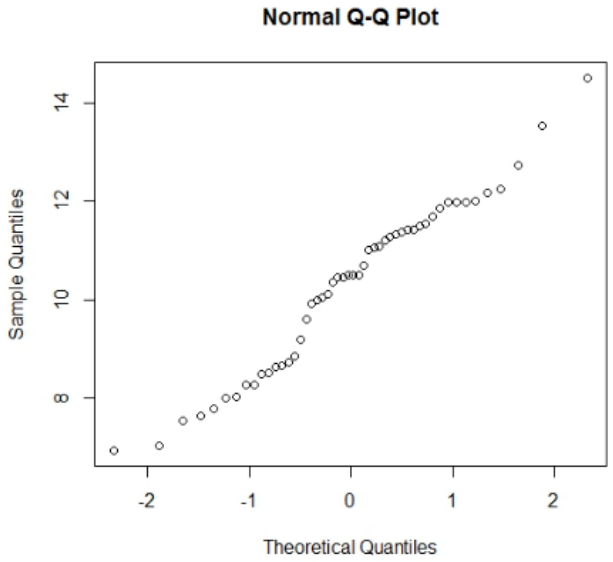
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

From my learnings, A **Quantile-Quantile (Q-Q) plot** is a graphical tool used to compare the distribution of residuals against a normal distribution. It helps check if a dataset follows a normal distribution. The Q-Q plot is a powerful tool that works for various sample sizes and helps detect distributional aspects such as shifts in location, scale, symmetry changes, and outliers. These insights are essential in validating the assumptions of linear regression

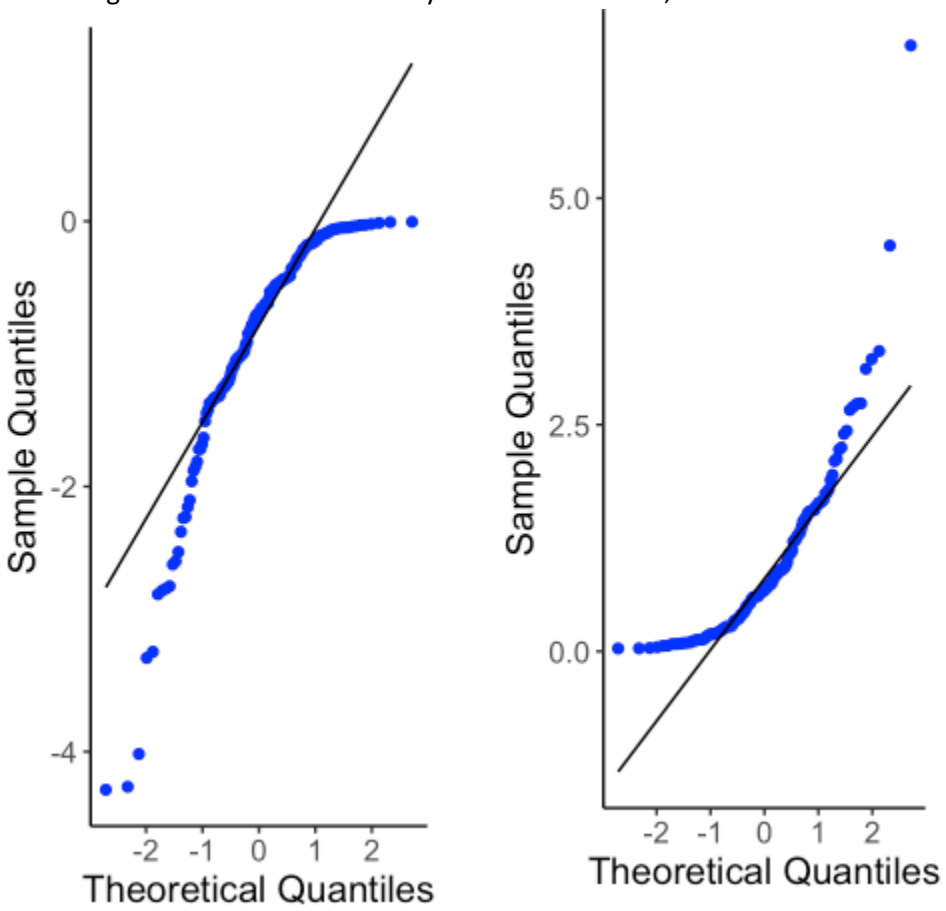
Q-Q Plot Importance in Linear Regression?

Linear Regression assumes that residuals (errors) are normally distributed. A Q-Q plot helps validate this assumption.

- If residuals are normally distributed, then points in the Q-Q plot will align along a 45° line.



- If residuals deviate from the line, it indicates skewness or non-normality. There left and right skewed distribution as you can see it below,



How to Interpret a Q-Q Plot?

1. Points lie on the 45° line - Normal distribution it means it's a good fit for Linear Regression.
2. Curved at ends - Indicates skewness it means the data is not symmetric.
3. S-shaped pattern - Heavy-tailed or light-tailed residuals it means there are potential outliers.

What If Residuals Are Not Normal?

1. Apply Transformations like Log, Square Root to make data normal.
2. Use Robust Regression, if there are outliers are present.
3. Switch to a Non-Linear Model if linear regression is inappropriate.

Conclusion: Q-Q plots are crucial in validating whether residuals are normally distributed, which ensures the reliability of p-values and confidence intervals in regression analysis.