upGrad | Bike Sharing Case Study | Linear Regression Model

Assignment-based Subjective Questions

Q.1 - From your analysis of the categorical variables, what can you infer about their effect on the dependent variable? (3 marks)

From the analysis of categorical variables in the bike-sharing dataset, several key inferences can be made about their effect on the dependent variable (cnt):

1. **Year (yr)**: Bike demand in 2019 is significantly higher by 1928.80 units compared to 2018, indicating an increasing trend in rentals.
2. **Season (season)**: Summer and winter seasons show higher bike demand by 806.27 and 1082.12 units, respectively, compared to spring, suggesting more favorable conditions or increased activities in these seasons.
3. **Month (mnth)**: Specific months such as March, August, September, and October have higher demand by 375.02, 649.41, 1060.85, and 460.11 units, respectively, compared to January, reflecting seasonal trends.
4. **Weekday (weekday)**: Saturdays have higher bike demand by 395.07 units compared to Sundays, likely due to leisure activities on weekends.
5. **Working Day (workingday)**: Higher bike demand on working days by 366.10 units suggests commuting patterns drive rentals.
6. **Weather Situation (weathersit)**: Misty or cloudy weather decreases bike demand by 450.38 units compared to clear weather, indicating that adverse conditions deter bike usage.

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

Using drop_first=True during dummy variable creation is important to avoid multicollinearity, which occurs when independent variables in a regression model are highly correlated. Multicollinearity can lead to difficulties in estimating the regression coefficients accurately and can inflate the variance of the coefficient estimates, making the model less reliable. By dropping the first category, we avoid the "dummy variable trap," ensuring that there is no redundancy among the dummy variables and the regression model remains well-defined and interpretable.

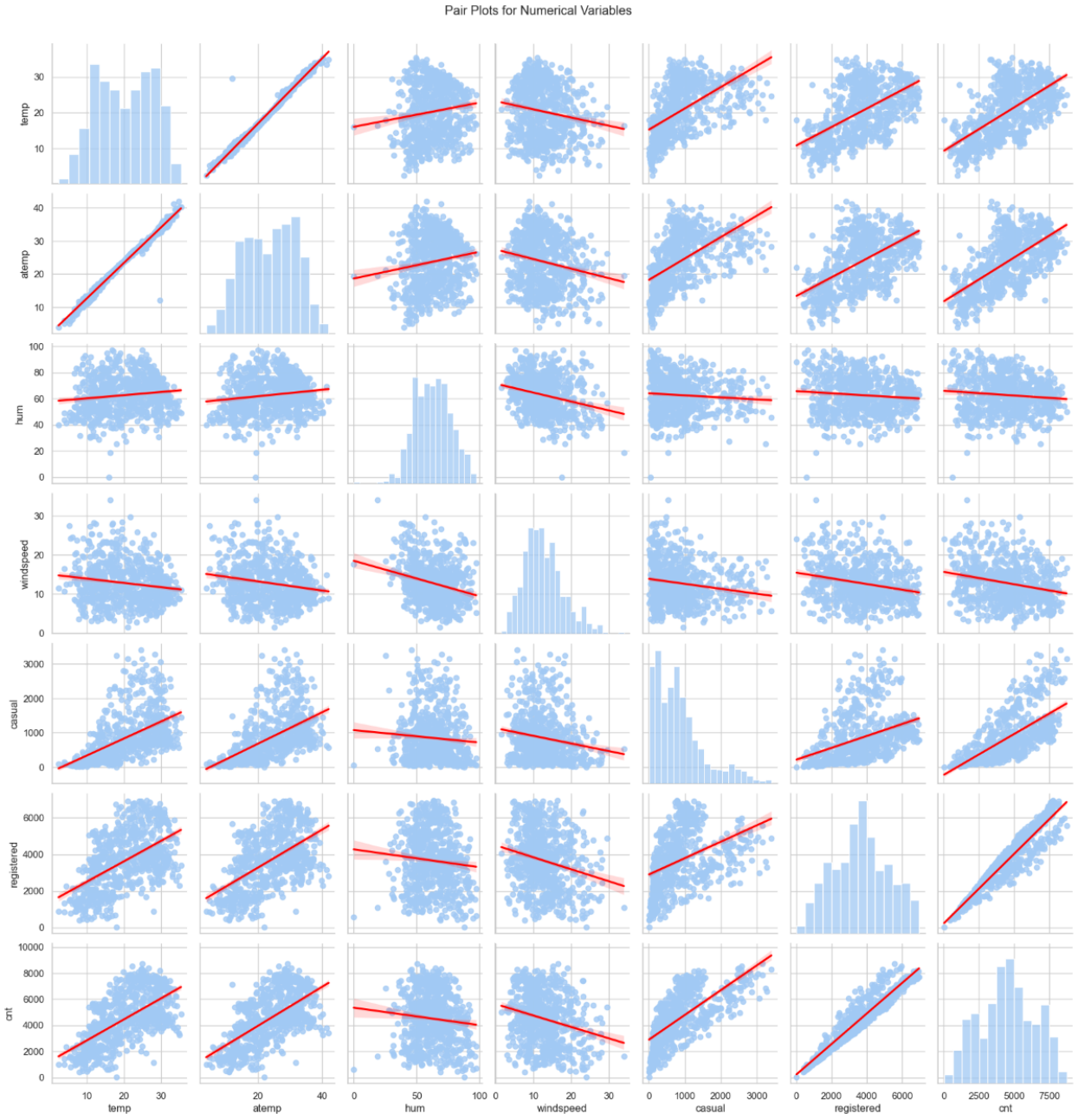
Here is typical code which implements drop_first when creating dummy variables using Pandas

```
data = pd.get_dummies(data, columns=categorical_vars, drop_first=True, dtype=int)
```

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

Looking at pair plot variable registered has the highest correlation with cnt.

Cnt is sum of casual and registered and registered has larger share in cnt, so this correlation is expected. However if I ignore casual and registered, variables temp and atemp have the highest correlation with cnt



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

To validate the assumptions of Linear Regression after building the model on the training set, the following checks were performed:

1. **Linearity and Homoscedasticity**: Residuals vs. fitted values plots for both training and test sets show that residuals are randomly scattered around the horizontal line, indicating no clear patterns. This suggests that the model's assumptions of linearity and homoscedasticity are reasonably met.
2. **Normality of Residuals**: Q-Q plots for both training and test residuals show that the residuals mostly follow the straight line, indicating that they are approximately normally distributed.
3. **Independence of Residuals**: The Durbin-Watson statistic of 2.02 suggests that there is no significant autocorrelation in the residuals.
4. **Multicollinearity**: Evaluated the VIF (Variance Inflation Factor) to ensure no predictor variables exhibited high multicollinearity.
5. The OLS summary was used to ensure all predictors are significant with very low p-values (< 0.05).
6. The model's performance metrics indicate strong predictive power with an Adjusted R Squared of 0.836, RMSE values of 739.44 (train) and 771.75 (test), and MAE values of 556.47 (train) and 587.49 (test), demonstrating consistent performance across both training and test datasets.
    - Training set R Squared: 0.84
    - Test set R Squared: 0.85
    - Adjusted R Squared: 0.836
    - RMSE values: 739.443 (train), 771.748 (test)
    - MAE values: 556.47 (train), 587.49 (test)

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

Based on the final model, the top 3 features contributing significantly towards explaining the demand for shared bikes are:

1. Temperature (temp) with a coefficient of 4164.17.
2. Year (yr) with a coefficient of 1928.80.
3. Humidity (hum) with a coefficient of -1197.26.

Winter season (1082.12) and wind speed (-1079.39) are righ next to humdidity in terms of impact.

General Subjective Questions

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

Linear regression is a supervised learning algorithm used to predict a continuous target variable based on one or more predictor variables.

The relationship between the target variable yy and the predictor variables XX is modeled by a linear equation:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$$

Here:

- y is the target variable.
- β_0 is the intercept.
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the predictor variables.
- $X_1, X_2, \ldots, X_n$ are the predictor variables.
- ϵ is the error term.

The goal of linear regression is to find the values of ββ (coefficients) that minimize the sum of squared residuals (the differences between the observed and predicted values). This is often achieved using the Ordinary Least Squares (OLS) method. The main steps involved are:

1. **Calculate the predicted values** using the initial or given coefficients.
2. **Compute the residuals** by subtracting the predicted values from the observed values.
3. **Minimize the sum of squared residuals** to find the best-fit line.

Advantages of Linear Regression Models

1. **Simplicity**: Easy to understand and implement.
2. **Interpretability**: Coefficients provide insights into the relationship between variables.
3. **Efficiency**: Computationally efficient for training and predictions.
4. **Scalability**: Can handle large datasets with linear complexity.
5. **Assumptions**: Works well if the relationship between the dependent and independent variables is approximately linear.

Disadvantages of Linear Regression Models

1. **Linearity Assumption**: Assumes a linear relationship, which might not hold true for all datasets.
2. **Outliers**: Sensitive to outliers, which can disproportionately affect the model.
3. **Multicollinearity**: Highly correlated independent variables can lead to unstable estimates.
4. **Homogeneity of Variance**: Assumes constant variance of errors, which might not be the case (heteroscedasticity).
5. **Limited Complexity**: Struggles with capturing complex relationships in the data.

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

Ref : https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Anscombe's quartet comprises four datasets that have nearly identical simple descriptive statistics yet appear very different when graphed. Each dataset has the same mean, variance, correlation, and linear regression line. The quartet demonstrates the importance of graphing data before analyzing it and the potential for statistics to mislead without visual inspection.

The four datasets as shown in image below:

1. Dataset I - A typical linear relationship.
2. Dataset II - A dataset where the relationship is nonlinear.
3. Dataset III - A dataset with an outlier that influences the regression line.
4. Dataset IV - A dataset where the x-values are all the same except one outlier.

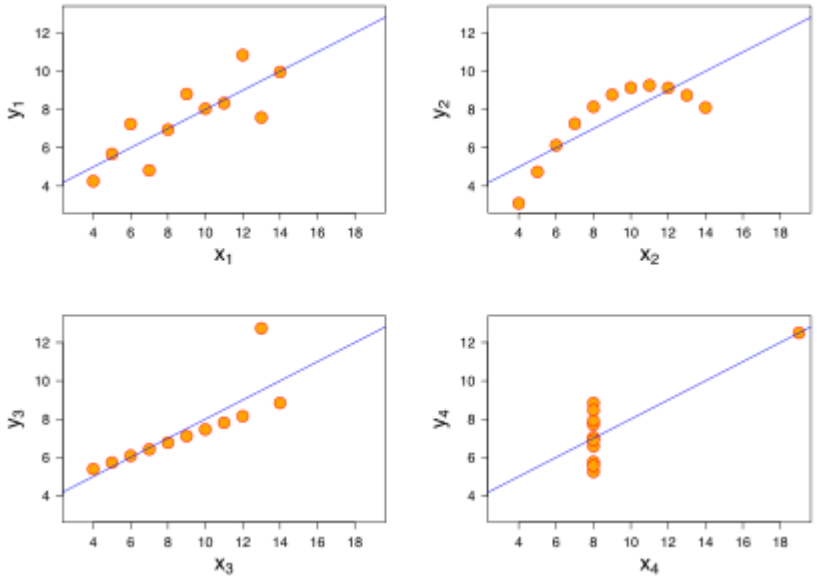
Anscombe's quartet

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

The four datasets have nearly identical simple descriptive statistics

Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x*: s_x^2	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y*: s_y^2	4.125	±0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression: R^2	0.67	to 2 decimal places

But when plotted each dataset has different story.

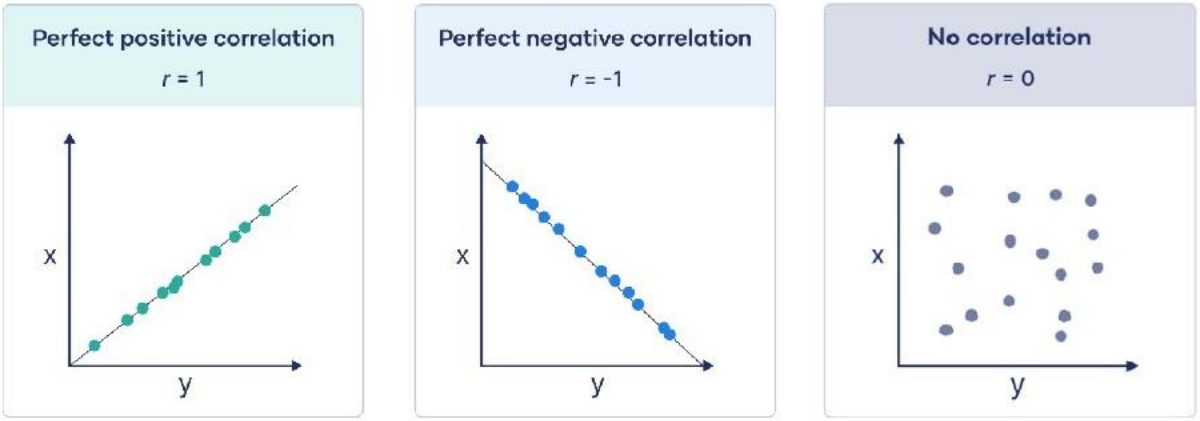


Q.3 - What is Pearson's R? (3 marks)

Ref : https://www.scribbr.com/statistics/pearson-correlation-coefficient/

Named after its inventor Karl Pearson, Pearson's R, also known as the Pearson correlation coefficient, measures the linear correlation between two variables. It ranges from -1 to +1, where:

- +1 indicates a perfect positive linear relationship.
- -1 indicates a perfect negative linear relationship.
- 0 indicates no linear relationship.



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

Ref : https://www.shiksha.com/online-courses/articles/normalization-and-standardization/

Scaling is the process of transforming the features of a dataset so that they have a similar scale, which can improve the performance and convergence of machine learning algorithms. While these terms are often used interchangeably, they serve different purposes and contexts.

Why Scaling is Performed:

- **Improves model performance**: Algorithms like gradient descent converge faster with scaled features.
- **Prevents dominance**: Features with larger ranges can dominate the learning process, leading to biased models.
- **Standardizes input**: Ensures that each feature contributes equally to the model.

Differences between Normalized and Standardized Scaling:

- **Normalization**:
- **Definition**: Scales data to a fixed range, typically 0 to 1.
- **Objective**: To change the scale of the variables so that they fit within a specific range.
- **When**: Use normalization when algorithms assume input features are on a similar scale or bounded range, such as neural networks, or to speed up gradient descent convergence, especially if data doesn't follow a Gaussian distribution or for models like k-nearest neighbors where variable magnitude is important.
- **Formula**:

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

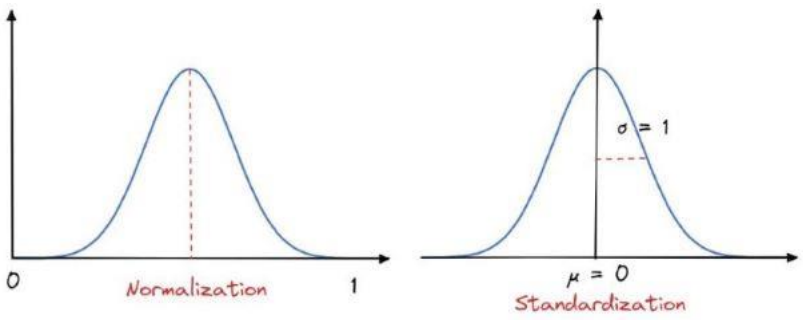
x' = Normalized Value
x = Original Value
$\min(x)$ = Minimum Value of x
$\max(x)$ = Maximum Value of x

- **Standardization**:
  - **Definition**: Transforms data to have a mean of 0 and a standard deviation of 1.
  - **Objective**: To change the distribution of the variables to a standard normal distribution.
  - **When**: Use standardization when algorithms assume input features are normally distributed with zero mean and unit variance, such as Support Vector Machines and Logistic Regression, or to handle outliers effectively.
  - **Formula**:

$$z = \frac{x - \mu}{\sigma}$$

z = Standardized Value
x = Original Value
μ = Mean of x
σ = Standard Deviation of x

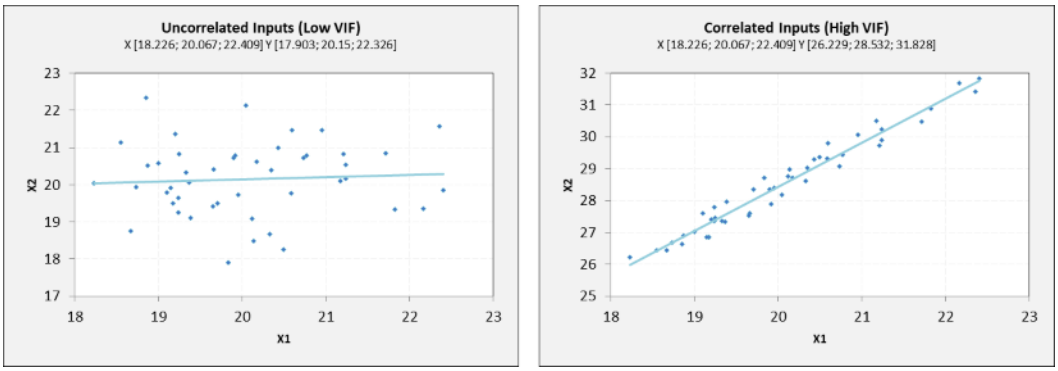
Here is typical graph of normalized vs standardized data



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

Ref : https://www.sigmamagic.com/blogs/what-is-variance-inflation-factor/

VIF (Variance Inflation Factor) measures the degree of multicollinearity among predictor variables in a regression model. An infinite VIF occurs when there is perfect multicollinearity, meaning that one predictor variable is a perfect linear combination of one or more other predictors. This situation makes it impossible to estimate the coefficients uniquely, leading to an infinite VIF.



Q.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 (Quantile-Quantile) plot is a graphical tool used to assess whether a dataset follows a given distribution, typically the normal distribution. It plots the quantiles of the sample data against the quantiles of a theoretical distribution. If the data follows the theoretical distribution, the points should form an approximate straight line.

Use and Importance in Linear Regression:

- **Normality Check**: It helps in checking the normality of residuals. Normally distributed residuals are an assumption of linear regression.
- **Identifying Outliers**: Deviations from the straight line indicate departures from normality, such as skewness or outliers.
- **Model Validation**: Ensures that the residuals meet the assumptions of the linear regression model, validating the model's reliability.

Q-Q plot used during model evaluation for Bike sharing assignment.