Assignment-based Subjective Questions

Analyzed by Sahadeb Patro

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

Below points illustrate the inferences about the effect of the categorical variables on the dependent variable (count of total rental bikes):

- **Year**: The demand for rental bikes increased significantly from 2018 to 2019, indicating a positive effect of year on the dependent variable.
- **Season**: The demand for rental bikes was highest in fall, followed by summer, winter, and spring. This suggests that the season has a strong effect on the dependent variable, with fall being the most favorable season for bike rentals.
- **Month**: The demand for bike rent varied throughout the months, with the highest demand in September and the lowest demand in January. This implies that month has a moderate effect on the dependent variable, with some months being more attractive for bike rentals than others.
- **Holiday**: The demand for rental bikes was slightly lower on holidays than on non-holidays, indicating a negative effect of holiday on the dependent variable.
- **Weekday**: The demand for bike rental was relatively consistent throughout the weekdays, with no significant differences between them. This suggests that weekdays have a weak effect on the dependent variable, with no clear preference for bike rentals on any day of the week.
- **Working Day**: The demand for bike rental was slightly higher on working days than on non-working days, indicating a positive effect of the working day on the dependent variable.
- **Weather Situation:** The demand for rental bikes was highest when the weather was clear, with few clouds or partly cloudy, followed by misty or cloudy, and lowest when there was light snow or rain. This implies that the weather situation has a strong effect on the dependent variable, with clear weather being the most conducive for bike rentals.

2. Why is it important to use drop_first=True during dummy variable creation?

drop_first=True is used to **convert categorical variables into dummy variables**. Dummy variables are **binary indicators** (0 or 1) that represent the presence or absence of a category. The reason why it is important to use drop_first=True is to **avoid the problem of multicollinearity**, which occurs when there is a high correlation among the dummy variables. Multicollinearity can cause issues in linear regression, by inflating the variance of the coefficients and making them unreliable.

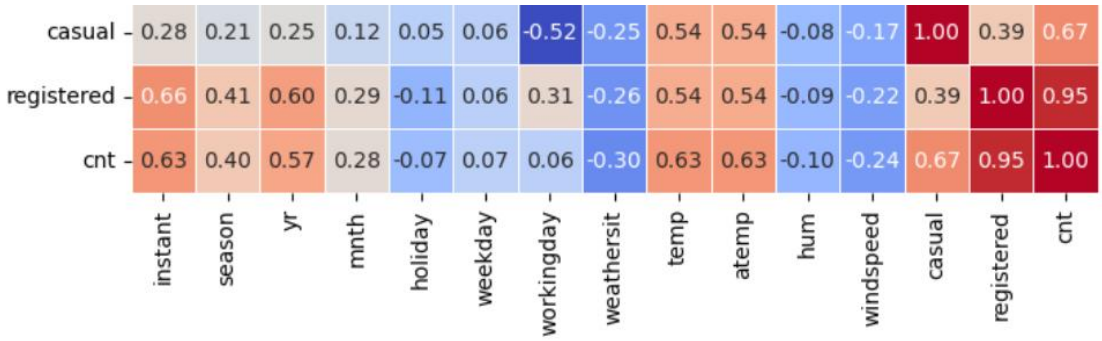
Some data science tools will only work when the input data is numeric. This is particularly true of machine learning. Many machine learning algorithms – like linear regression and logistic regression – strictly require numeric input data. If you try to use them with string-based

categorical data, they will throw an error. So, before you use such tools, you need to encode your categorical data as numeric dummy variables. This is where using pandas property drop_first=True come in place.

Gender	Gender_male	Gender_female
M	1	0
M	1	0
F	0	1
F	0	1

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

Initially, **Registered** columns which has highest level of correlation with the target variable followed by **Casual**.



However, after performing a few data cleaning activities **temp** column became highly correlated with target variable.

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

To check the **linearity assumption**, I plotted the scatter plots of the dependent variable (count) with each independent variable and looked for any nonlinear patterns. It also calculated the correlation coefficients between the variables and looked for any values close to zero or one.

To check the **multicollinearity assumption**, I calculated the VIF for most independent variables and looked for any values greater than 10, which indicate high multicollinearity. It also plotted the correlation matrix of the independent variables and looked for any values close to one or negative one, which indicate strong positive or negative correlation.

I also looked for any **p-values less than 0.05**, which indicate significant departure from normality.

Assumption of Error terms is independent: we see that there is almost no relation between Residual and predicted value.

To check the **homoscedasticity assumption**, I plotted the residuals versus the fitted values and looked for any patterns or trends that indicate unequal variance. I noticed for any p-values less than 0.05, which indicate heteroscedasticity.

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

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

Temperature in Celsius: This feature has a positive coefficient of 0.47, indicating that higher temperatures are associated with higher demand for shared bikes.

Year: This feature has a positive coefficient of 0.23, indicating that the demand for shared bikes increased from 2018 to 2019.

Weather Situation - Snow (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds): This feature has a negative coefficient of -0.29, indicating that snowy weather conditions are associated with lower demand for shared bikes.

General Subjective Questions:

1. **Explain the linear regression algorithm in detail.**

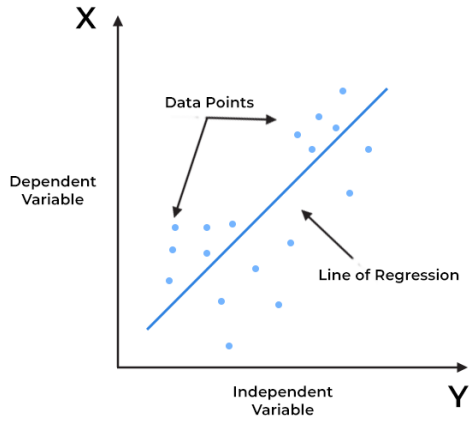
In a simple term, Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. The goal is to find the best-fitting line (or hyperplane in higher dimensions) that minimizes the difference between the predicted values and the actual values. It helps us understand how one thing changes when other values change.

Process for linear regression:

1. **Identify the dependent variable** (also called the target or response variable) that you want to predict. Also, Identify one or more independent variables (also called features or predictors) that you believe influence the dependent variable.
2. **Data Sourcing**: Collect dataset that includes values of both the dependent and independent variables.
3. Perform **EDA** to understand relationships between the variables using graphs and statistical measures. Understand the general trends and patterns in the data.
4. **Divide the dataset** into two - Test & Train

5. **Create a Model**:
    a. In **simple linear regression**, there is one independent variable, and the relationship is modeled using a straight line **(y = mx + b).**
    b. In **multiple linear regression**, there are multiple independent variables, and the relationship is modeled using a hyperplane in higher dimensions.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$



6. **Train and Test** Model with respective data.
    a. Adjust the coefficients iteratively to minimize the objective function. This is typically done using optimization algorithms.
    b. The model "learns" from the data, adjusting its parameters to make better predictions.
7. Use models and make predictions:
    a. Once the model is trained, you can use it to make predictions on different datasets.

Key Points:

1. Linear regression helps us find a simple relationship between variables.
2. It's like drawing a line through points to predict values.
3. The goal is to minimize the difference between predicted and actual values.
4. Once trained, the model can make predictions for new data.

2. Explain the Anscombe's quartet in detail.

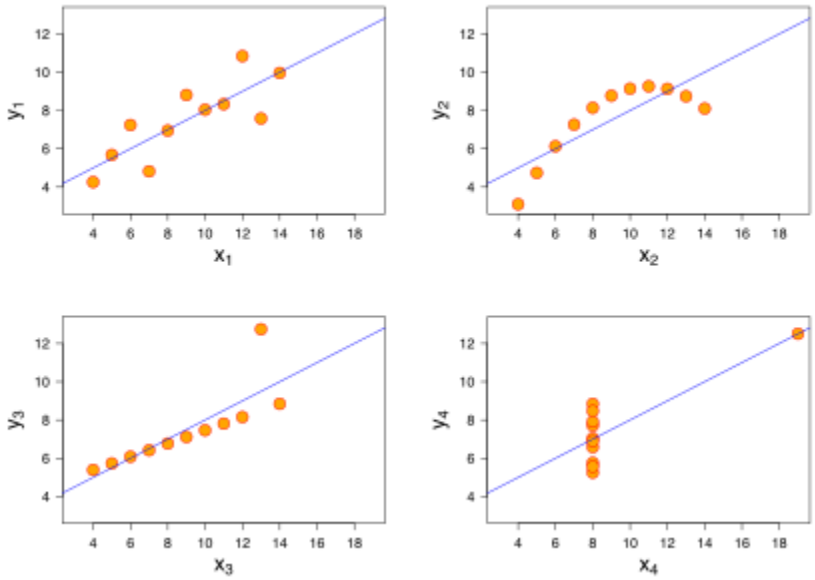
Anscombe's quartet is a **set of four datasets that have the same mean, standard deviation, and correlation for both x and y variables**, **but look very different when plotted on a graph**.

The quartet was created by the statistician Francis Anscombe in 1973 to **show the importance of visualizing data before analyzing it**, and to demonstrate that summary statistics alone can be misleading.

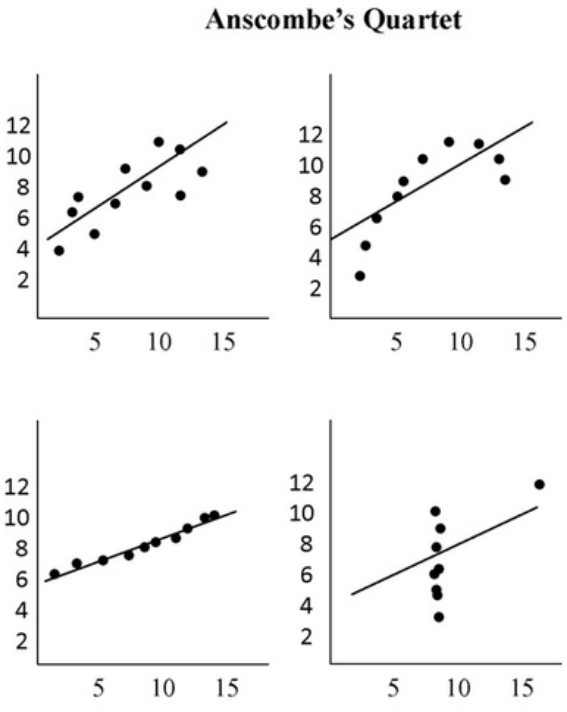
Each dataset consists of 11 (x, y) points. The first dataset (top left) looks like a simple linear relationship, where y increases as x increases. The second dataset (top right) has a curved

relationship, where y follows a parabolic pattern. The third dataset (bottom left) also has a linear relationship, but with one outlier that affects the slope and correlation. The fourth dataset (bottom right) has no relationship at all, except for one point that creates a high correlation.

The quartet teaches us that we should always plot our data and look for patterns, outliers, and other features that might not be captured by summary statistics. It also reminds us that correlation does not imply causation, and that we should use appropriate methods to test our hypotheses and model our data.



Example:



Data sets for the 4 XY plots

I		II		III		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	5.76
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	8.81	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	7.26	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

All the summary statistics for each dataset are the same:

- The mean value of x is 9.
- The mean value of y is 7.5.
- The variance for x is 11, and for y, it is 4.12.
- The correlation between x and y is 0.816.
- The best-fit line is y = 0.5x + 3.

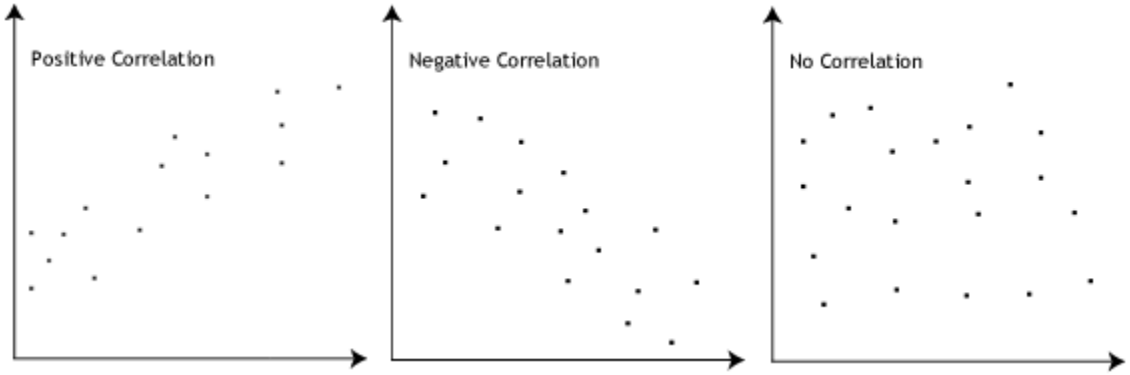
However, the visual representations convey distinct narratives for each dataset. The initial scatter plot (top left) suggests a straightforward linear relationship, indicating a correlation between two variables where y might follow a Gaussian distribution with the mean linearly dependent on x.

Why this is important:
Anscombe's quartet shows us that only looking at summary numbers can be misleading. Even if averages and correlations seem similar, the actual patterns in the data might be very different. This underlines how crucial it is to use graphs and charts to really grasp the relationships in the data, and not to make conclusions based solely on numbers.

3. What is Pearson's R?

Pearson's R, or just "R," is a number that tells us how much two things are connected in a straight line.

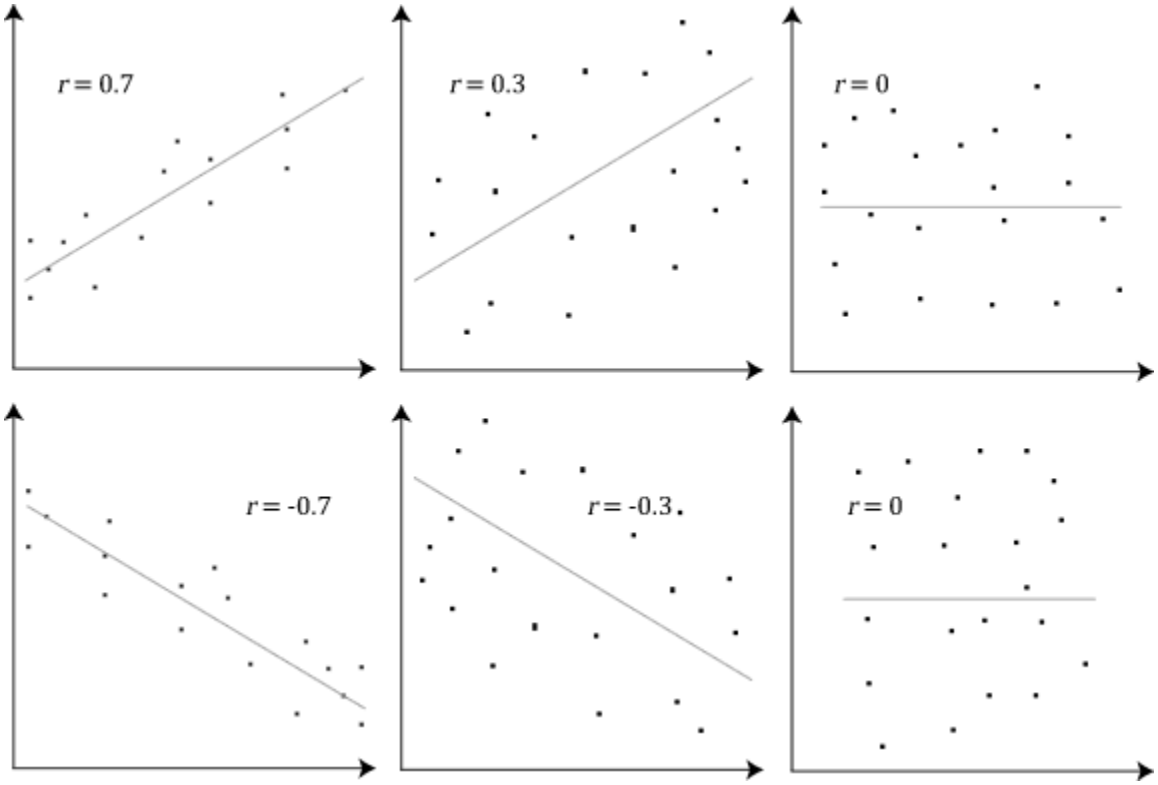
- If R is close to 1, it means they're strongly connected in a positive way (both go up together).
- If R is close to -1, it means they're strongly connected but in a negative way (one goes up while the other goes down).
- If R is close to 0, it means there isn't much of a straight-line connection between them.



The Pearson correlation coefficient, denoted as "r," tends to approach either +1 or -1 when there is a strong association between two variables. A positive value of +1 indicates a positive relationship, while a negative value of -1 suggests a negative relationship. In such cases, all data points align closely with the best-fit line, displaying minimal variation.

Correlation coefficients falling between +1 and -1, such as 0.8 or -0.4, signify that there is some

variation around the best-fit line. The closer the coefficient is to 0, the greater the variation observed around the line of best fit. The diagram below illustrates different relationships and their corresponding correlation coefficients.



So, R helps us understand how changes in one thing are linked to changes in another thing, and the number itself gives us a clue about how strong that link is.

Formula:

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

Where,
r = Pearson correlation coefficient
x = Values in the first set of data
y = Values in the second set of data
n = Total number of values.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

What is Scaling?

Scaling is a process in data preprocessing that involves adjusting the range of values in a dataset to ensure that all features contribute equally to the analysis. It is **performed to bring**

numerical features to a similar scale, preventing certain features from dominating due to their larger magnitudes. There are different methods for scaling data, with the goal of making features comparable and enhancing the performance of machine learning algorithms.
Why is Scaling Performed?

- Scaling is **performed in data preprocessing to ensure equal contribution** of all features, preventing larger-scale features from disproportionately influencing the model.
- It accelerates convergence speed in **optimization algorithms, improves performance** in distance-based algorithms, enhances regularization effectiveness, and maintains consistency in interpreting coefficients in linear models.
- This step creates a level playing field for features, contributing to more effective machine learning models.

Normalized scaling and standardized scaling are two common scaling techniques:

Normalized Scaling:

- Also known as Min-Max scaling.
- Rescales the values to a specific range, usually between 0 and 1.
- Formula:

$$\text{Normalized Value} = \frac{\text{Original Value} - \text{Min}}{\text{Max} - \text{Min}}$$

- Useful when the features have different ranges and you want them to be within a consistent scale.

Standardized Scaling:

- Also known as Z-score normalization.
- Transforms the data to have a mean of 0 and a standard deviation of 1.
- Formula:

$$\text{Standardized Value} = \frac{\text{Original Value} - \text{Mean}}{\text{Standard Deviation}}$$

- Useful when the features have different units or follow different distributions, ensuring they are comparable.

Normalized scaling and standardized scaling are two common techniques, each serving different purposes in bringing features to a consistent scale.

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

VIF measure of how much the variance of a regression coefficient is inflated due to multicollinearity among the independent variables.

Multicollinearity occurs when some of the independent variables are highly correlated with each other, which means they can explain each other's variation. This makes it difficult to estimate the unique effect of each variable on the dependent variable.

The formula for VIF is:

$$VIF = \frac{1}{1 - R^2}$$

where R2 is the coefficient of determination for the regression of one independent variable on the rest of the independent variables.

An infinite value of VIF for a given independent variable indicates that it can be perfectly predicted by other variables in the model. Looking at the equation above, this happens when R2 approaches.

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

This means that there is a perfect linear relationship among the independent variables, which violates the assumption of no multicollinearity for a valid regression analysis.

To avoid this problem, we need to identify and remove the variables that cause multicollinearity, or use other methods such as regularization, principal component analysis, or ridge regression.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**

A Q-Q (Quantile-Quantile) plot is a visual tool used to evaluate whether a dataset is likely to follow a specific theoretical distribution, such as the Normal, exponential, or Uniform distribution. It is also employed to compare two datasets and determine if they share a common distribution.

In the context of linear regression, where we may have separate training and test datasets, the Q-Q plot becomes crucial. It helps confirm whether both datasets originate from populations with the same distributions. Some advantages include its applicability to various sample sizes and its ability to detect shifts in location, scale, changes in symmetry, and the presence of outliers.

The Q-Q plot is useful for checking scenarios such as whether two datasets:

- Originate from populations with a common distribution,
- Share common location and scale,
- Exhibit similar distributional shapes,
- Demonstrate similar tail behavior.

A Q-Q plot is constructed by plotting quantiles from two sets against each other. If the quantiles from both sets align, it suggests that they are derived from the same distribution.