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?

ANSWER: following are some of the inferences about the effect of categorical variables on dependent variable

1. if we look at seasons most number of bike rentals are in fall and then followed by summer

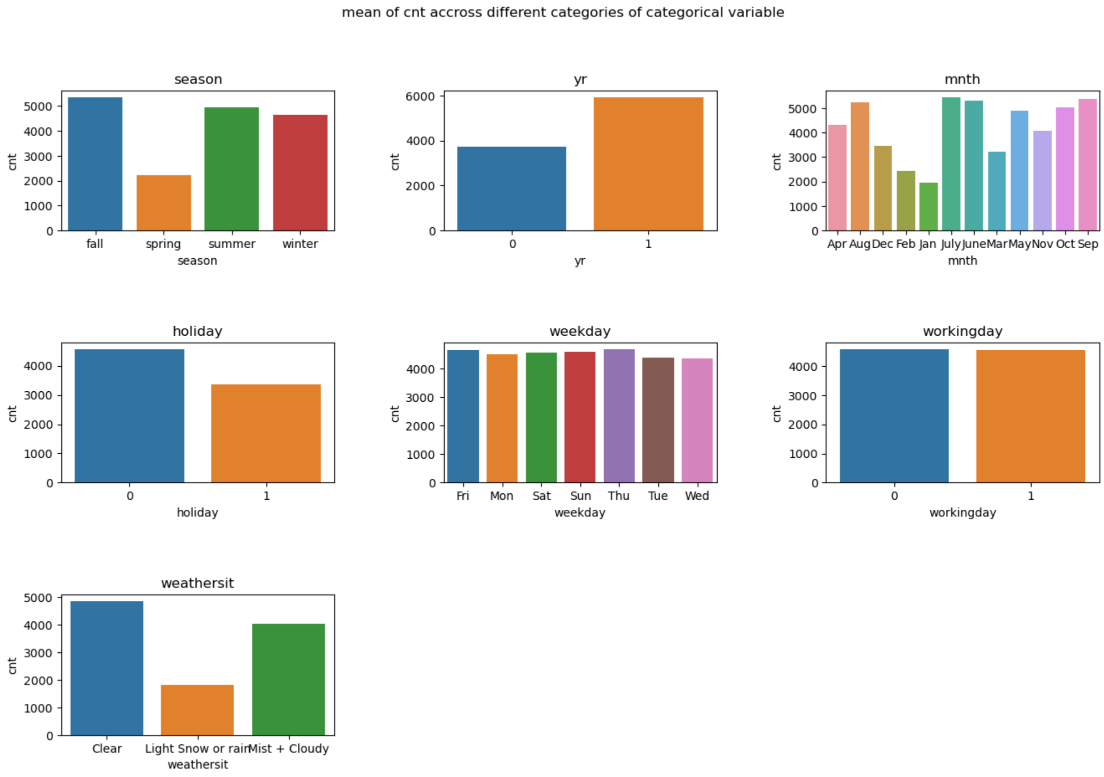
2. there is significant increase in the bike rentals from 2018 to 2019 so we can say that the deman for bike rentals are increasing rapidly

3. bike rentals are very high on non-holiday days compaired with holiday this may be because most of the people are using rental bikes for job and daily activities

4. among all the week days friday and thusday have the highest bike rentals on average

5. if we look at working day variable alomst both working day and non working day are almost same .

6. its very obvious that there will be some kind of effect of weather on the customers which is also conformed by the weather situation plot where we can see that the deman for bike rentals are considerable very high in clear weather followed mist and cloudy and very low during rainy or snow days.


mean of cnt accross different categories of categorical variable

this plots are from submitted jupyter notebook

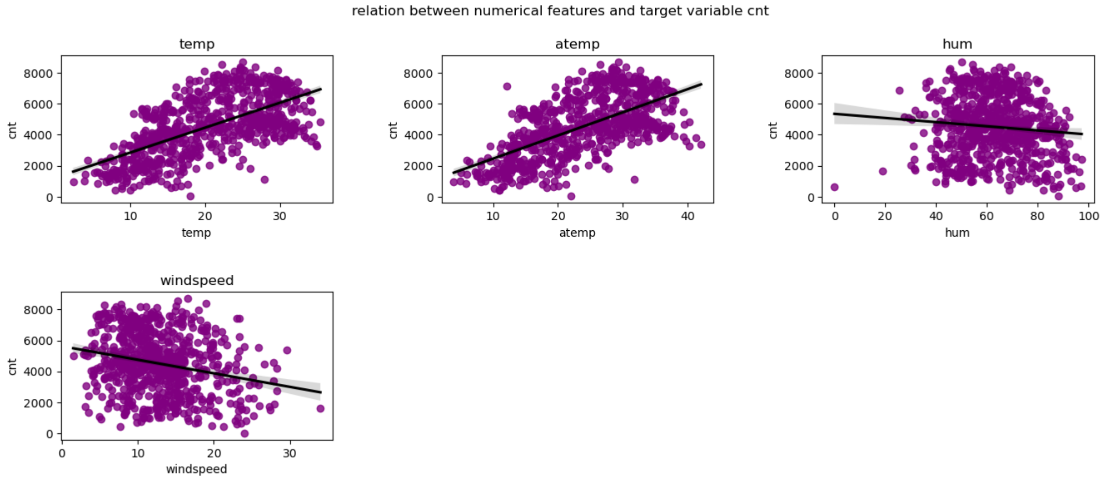
2.Why is it important to use drop_first=True during dummy variable creation?

ANSWER: Dummy variables are created to represent categorical variables as numerical values in regression models. if there is a categorical variables with 'n' number of categories then we only need n-1 dummy variables because we can represent the 'n'th variable with the n-1 variables.

When creating dummy variables from categorical variable in a regression model, setting drop_first=True is a common practice that helps to drop one of the dummy variables created so that we can end up with n-1 dummy variables this also helps to avoid multicollinearity and improves the interpretability of the model. Multicollinearity occurs when two or more predictor variables are highly correlated, which can lead to unstable and unreliable coefficient estimates in the regression model.

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

ANSWER: from the below pair-plot we can see that **temp** has the highest correlation with the target variable which is **0.63** higher than all other features



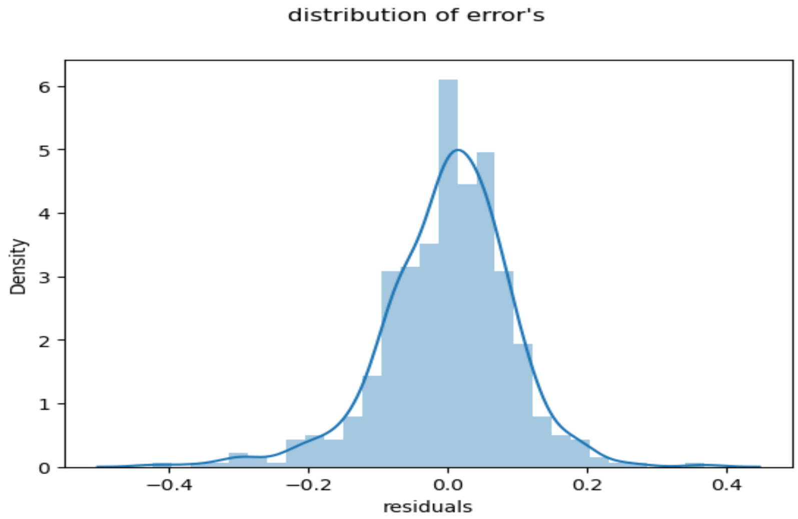
this plots are from submitted jupyter notebook

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

ANSWER: if we just want to find the best fit line then the only assumption made is that the target variable and the input variables are linearly dependent. but if we want to inferences sample to population which we generally do then we need to make some assumptions.

1. Error terms are normally distributed with mean zero :

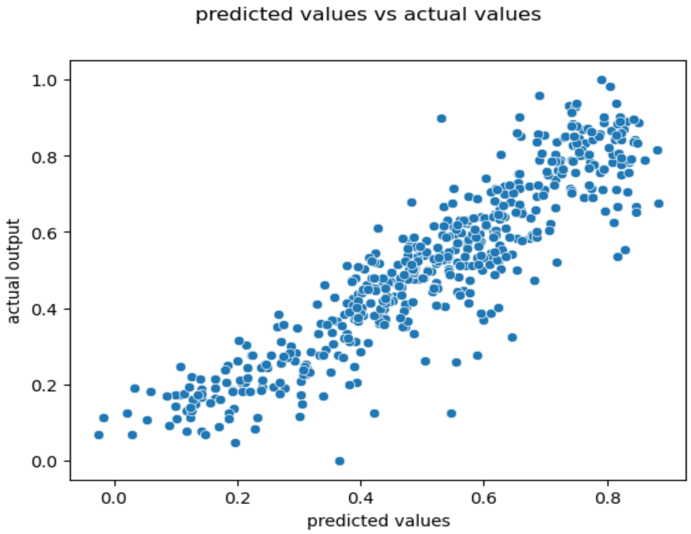
this assumption can be validated by ploting a distribution plot of residuals as shown below


distribution of error's

from the above dist plot we can see that residuals are normally distributed with mean zero hence we can say that this assumption is validated

2. Error terms have constant variance (homoscedasticity):

In linear regression, the assumption of homoscedasticity, or constant variance of error terms, is crucial for the reliability of the model's predictions. When this assumption holds true, it means that the spread of the residuals (the differences between observed and predicted values) remains roughly consistent across all levels of the independent variables.


predicted values vs actual values

this plots are from submitted jupyter notebook we can validate this assumption by ploting the sactter plot between predicted values and actual values .we essentially comparing the model's predicted outcomes with the real observed outcomes

In the context of heteroscedasticity, we are looking for patterns in the spread or dispersion of the data points along the y-axis as the x-axis values change.

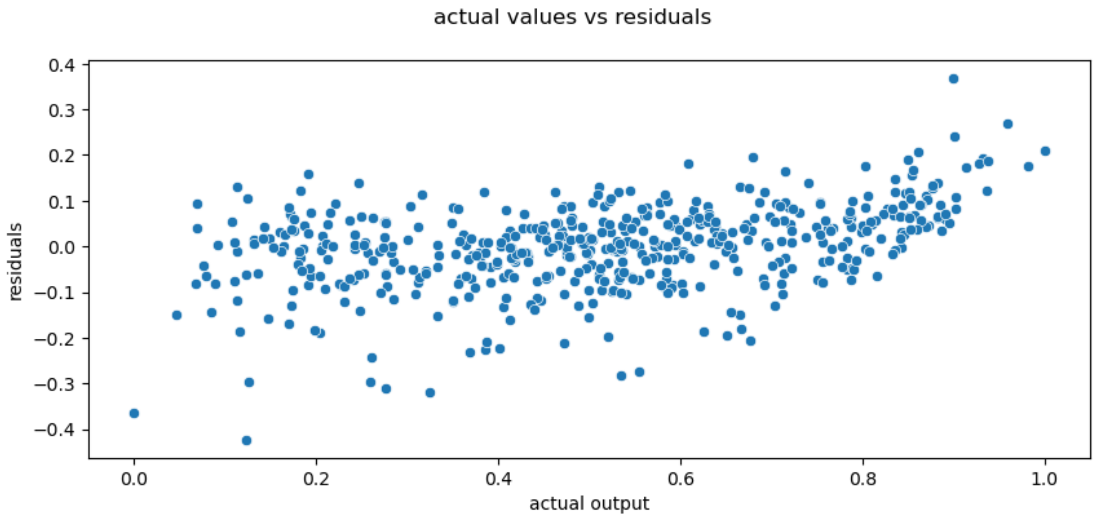
If the spread of data points along the y-axis is consistent across all levels of predicted values on the x-axis, we can say that the assumption of homoscedasticity is validated.

3. Error terms are independent of each other:

this assumption states that residuals should not follow any perticular pattern in other words pattern of errors should not provide information about the errors they should be complitely random .

we can validate this assumption using a scaltter plot of residuals vs actual output values. in this plot there should not be any Systematic Pattern,No Fan or Funnel Shape

The spread of the residuals along the horizontal axis should be fairly consistent across the range of actual output values.then we can say that this assumption is validated



this plots are from submitted jupyter notebook

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

ANSWER: following are the top 3 variables contributing significantly towards explaining the deman

1. temp(0.44)

2. Light Snow or rain(-0.285463)

3. yr (0.234244)

General Subjective Question

1.Explain the linear regression algorithm in detail.

ANSWER:

- Linear regression is a fundamental statistical technique used for modeling the relationship between a dependent variable and one or more independent variables.
- the goal of the linear regression is to find the best-fitting linear equation that describes the relationship between these variables.
- linear regression is a supervised machine learning algorithm which takes one or more inputs and find the equation for best fitting line
- Mathematical Representation of the linear regression equation is

  y = c + (m1*x1)+(m2*x2)+(m3*x3)+......+(mn*xn)+e

  where :

  y = is the dependent variable

  x1,x2,x3...xn = are the independent variables

  m1,m2,m3...mn = are the coefficients (also known as regression coefficients or weights).

  e = represents the error term (the difference between the actual 'y' and the predicted 'y' by the model).

- The goal is to determine the coefficients that minimize the sum of squared differences between the actual 'y' values and the predicted 'y' values (residuals). This process is usually achieved through a method called Ordinary Least Squares (OLS).
- The algorithm iteratively adjusts the coefficients to minimize the sum of squared residuals. It finds the line that best fits the data points in a way that the vertical distance between each data point and the line (the residual) is minimized.
- Linear regression relies on certain assumptions such as linearity (the relationship between variables is linear), independence of errors, homoscedasticity (constant variance of errors), and normally distributed errors,errors should have mean zero.
- specially in case of multi-linear regression model we need to additionally consider Overfitting,Multicollinearity,Feature selection
- after training the model then we can evaluate the model using R-squared value which tell how well the model explains the variance in the dependent variable. for eg if R-squared value is 0.80 then it means 80% of the variance is explained by the model.
- The trained model can be used to make predictions for new or unseen data by plugging in the values of the independent variables into the regression equation.

2. Explain the Anscombe’s quartet in detail.

- Anscombe's quartet demonstrates the need for data visualization to understand the underlying relationships, identify outliers, and make informed decisions. It emphasizes that relying solely on summary statistics can lead to incomplete insights and misinterpretations of data.

- Anscombe's quartet is a collection of four datasets that have nearly identical statistical properties, but they exhibit remarkably different patterns when plotted and analyzed. These datasets were created by the statistician Francis Anscombe in 1973 to illustrate the importance of visualizing data and the limitations of relying solely on summary statistics.

- Anscombe's quartet underscores the idea that visual exploration is crucial for understanding data and detecting potential relationships that might not be apparent from statistics alone.

- Dataset 1:

  x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7

  y: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82

- Dataset 2:

  x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7

  y: 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26

- Dataset 3:

  x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7

  y: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42

- Dataset 4:

  x: 8, 8, 8, 8, 8, 8, 8, 19, 8, 8

  y: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91

- upon perform some basic statistical analysis on each of the datasets in Anscombe's quartet we will get the below statistics:

- Mean of x: 9.0

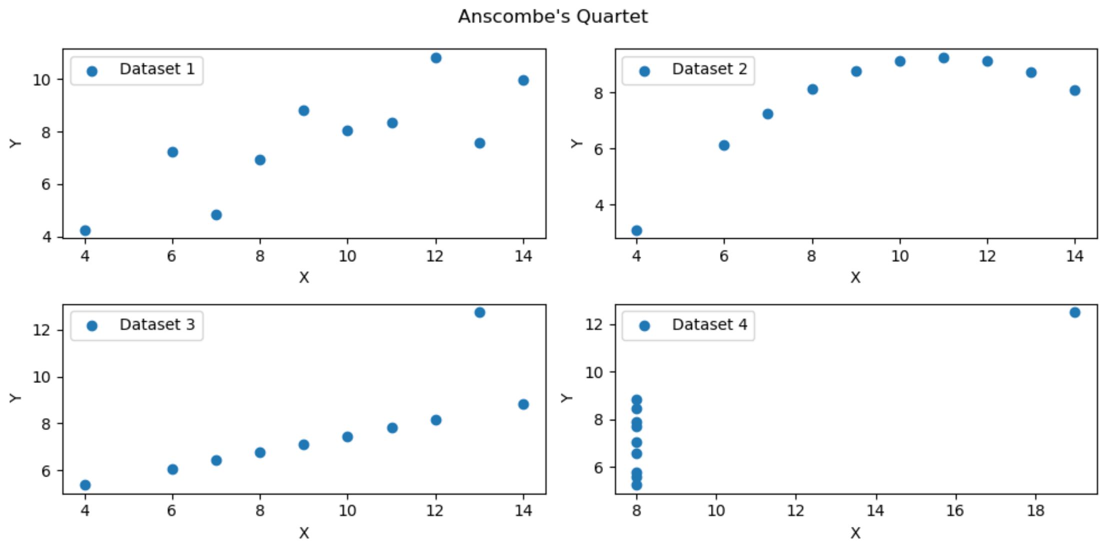
- Mean of y: 7.50

- Variance of x: 11.0

- Variance of y: 4.125

- Correlation between x and y: 0.816

- Despite having similar summary statistics (means, variances, correlations), these datasets reveal different patterns when visualized. For instance:



this plot is ploted in the jupyter notebook

- Dataset 1 and 2: These datasets seem to exhibit a linear relationship between x and y, suggesting a linear regression model could fit well.

- Dataset 3: This dataset showcases an outlier that significantly influences the linear regression fit. Removing the outlier would result in a much better linear fit.

- Dataset 4: Here, the presence of an outlier significantly impacts the correlation coefficient, but a linear regression line does not capture the overall pattern of the data.

3. What is Pearson’s R?

- pearson's R is used to measure the strength and direction of a linear relationship between two continuous variables

- It ranges from -1 to +1

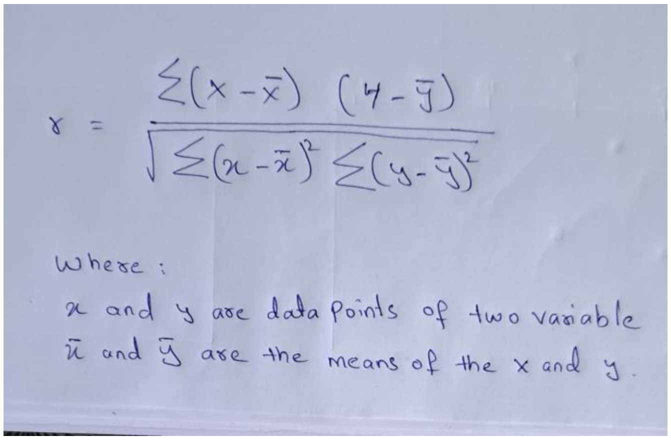
- where:

  R=+1 indicates a perfect positive linear correlation (as one variable increases, the other also increases proportionally).

  R=-1 indicates a perfect negative linear correlation (as one variable increases, the other decreases proportionally).

  R=0 indicates no linear correlation.

- formula for Pearson’s R is



$$r = \frac{\sum(x-\bar{x})\ (y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

where :

x and y are data points of two variable

$\bar{x}$ and $\bar{y}$ are the means of the x and y .

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

- scaling is the process of transforming numerical variables to a standardized range
- Scaling is essential because it ensures that different variables with different units or ranges contribute equally to the analysis and modeling processes
- in real world data may consists of many variables with different range of values such as one variable with (100-1000) range and another variable of range (10-100) in such cases one variable may dominate the other as a result Algorithms may give more importance to the variable with large range although both the variables are of equal importance
- Also In optimization algorithms like gradient descent, features with larger scales can lead to uneven convergence.
- Due to the above mentioned resoan we perform scaling before start building the model
- there are mainly two common scaling techniques they are Min-Max Scaling and Standardization

1. Min-Max Scaling :

   - Transforms features to a common range, making them comparable

   - Useful when algorithms are sensitive to feature scales, like gradient descent

   - Sensitive to outliers

   - min-max scales the variable such that its max value is 1 and minimum value is 0

2. Standardization:

   - Doesn't get affected by outliers as much as min-max scaling.

   - Suitable when the data distribution is not necessarily normal.

   - Can result in negative values

- Doesn't make all features fall within a specific range rather values are concentrated arround mean with a unit standard deviation

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

- VIF stands for Variance Inflation Factor which is used to measure the multicollinearity between independent variables in a linear regression model
- A higher VIF indicates higher multicollinearity
- to undersatand the scenario when VIF is infinity we need to look at its formula

VIF = 1/(1-R^2)

- lets say we have two variables A and B where B is the compliment of A so they are negativly correlated hence if we calculate tehe VIF of this model then A or B will get VIF as infinity
- when two variables are exactly correlated then R^2 will be come 1 and when R^2 becomes 1 then in the denominator we will get 1-1 = 0 hence will lead to the VIF to infinity
- VIF becomes infinite can occur due to perfect multicollinearity.
- Perfect multicollinearity happens when two or more independent variables are perfectly correlated means that one variable can be perfectly predicted from the others. In such cases the regression coefficients are not uniquely determined

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

- Q-Q plot stands for Quantile-Quantile plot
- it is use to check the similarities between the practical sample data and theoretical distribution
- for example residuals should be normally distributed when talking about theoretical distribution. Q-Q plots are often used to check whether the residuals of a regression model follow a normal distribution.
- here is how Q-Q plot works:
- Sorting Mistakes:First you look at all the mistakes your model made when predicting something. You organize them from the smallest mistake to the biggest mistake.
- Seeing Where They Stand:You want to see how each mistake compares to the others So you figure out how many mistakes are smaller than the smallest one how many are smaller than the second smallest and so on This helps you understand where each mistake stands among all the mistakes.
- Comparing to a Normal Picture:Now you imagine what these mistakes would look like if they followed a common pattern (like a normal distribution) You expect some mistakes to be very small some to be medium-sized, and some to be big just like how things often happen in real life.
- Drawing the Plot:You create a plot with two lines One line shows where the mistakes would stand if they perfectly followed the expected pattern. The other line shows where your actual mistakes are standing.
- If most of your mistakes fall close to the line it means your mistakes are quite normal just like expected.

- If the mistakes go away from the line it suggests that they're not following the pattern you expected This could mean something is not quite right with your model's predictions.
- In short a Q-Q plot helps you see if the mistakes your model makes are normal or not It's like comparing your real-life mistakes to what you would expect to happen by chance. If they match great If not, you might need to look more closely at your model's behavior.

name : Gariganti Rahul