Bike Sharing Assignment

Subjective questions response by Parijaat Sunil

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: The effect of the categorical variable is given below

a. **Year** – The number of users is growing year on year, and hence each consecutive year has a positive effect on the number of users.
b. **Month** – The months of June to August have the highest number of users. However, the model seems to show a small negative effect of July month and a smaller positive correlation with September month. This may mark the start and end of the rainy periods when most users would not be using bikes
c. **Weekday** – Weekdays do not seem to have a major effect on the dependent variable.
d. **Holiday** – Holidays seem to have a negative effect on the dependent variable. This may be due to the fact that many users use bikes to commute to work.
e. **Season** – The spring season seems to have a negative effect on the dependent variable while winter has a low positive effect on the dependent variable. Fall is the month with the highest number of users.
f. **Weather** – Light rain has a negative effect on the dependent variable. The highest number of users are found on days with clear weather.

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

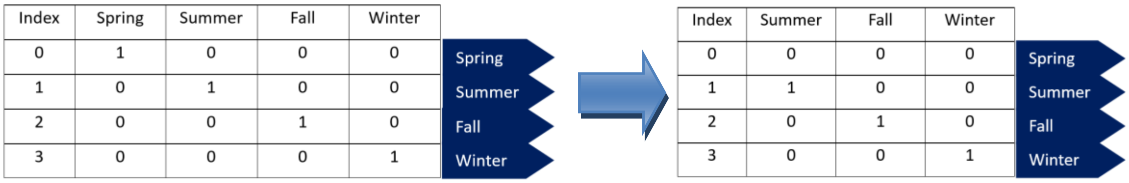
Answer: We use dummy variables to convert categorical variables into a numerical format which can be used by the linear regression model to measure the effect of each category on the dependent variable.

The pd.get_dummies method takes all the available unique categories in the applied column, and converts them to columns with binary format to show which category each data row belongs. To denote a column containing ‘n’ categories, we would need ‘n-1’ rows, as a 0 or false in ‘n-1’ columns means that the data row belongs to the ‘n’th category.

The drop_first parameter drops the first column in the dataset created by the pd.get_dummies method. This is done to keep the model simpler by reducing the number of redundant features.

The below example using the Season feature explains the above concept:

Without using the drop_first parameter: With the drop_first=True parameter:

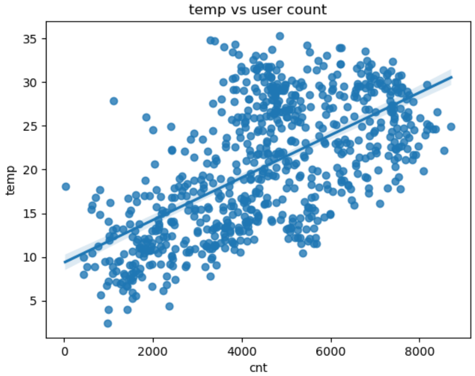


Index	Spring	Summer	Fall	Winter	
0	1	0	0	0	Spring
1	0	1	0	0	Summer
2	0	0	1	0	Fall
3	0	0	0	1	Winter

Index	Summer	Fall	Winter	
0	0	0	0	Spring
1	1	0	0	Summer
2	0	1	0	Fall
3	0	0	1	Winter

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

   Answer: From the pair plot, we can see that the temp feature has the highest correlation with the target variable. Both temp and atemp have a similar and the highest correlations with the target variable. The scatter plot of the



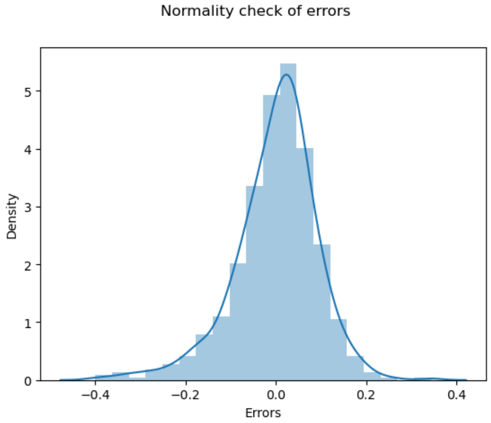
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

   Answer: The created model was validated against the below assumptions of Linear Regression

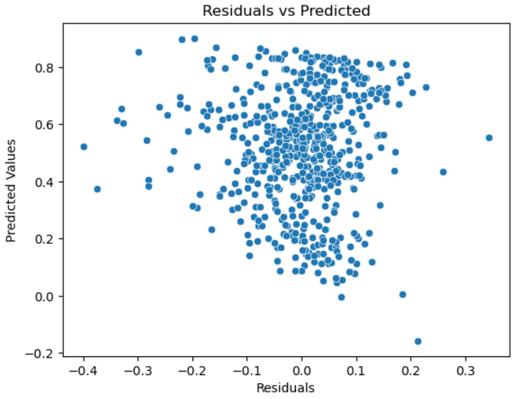
   a. **Linearity**: We checked that there is a linear relationship between the predictors and independent variables. This is shown by the high Rsquared and very low p value of f-static.

   b. **Multicollinearity**: We checked the Variable Inflation Factor (VIF) for all predictors after each developed model. The value of VIF was found to be lesser than 5 in the final model, which was acceptable.

   c. **Normality**: We plotted the errors and found them to be following a normal distribution with the mean centered around 0. The graph is provided below:



Check for Normality



Check for homoscedasticity

   d. **Homoscedasticity**: We plotted the residual values and found that there is no visible pattern, meaning that the variance of errors is constant. The graph is provided above.

   e. **Autocorrelation**: The Durbin – Watson (DW) statistic for the developed final model is 1.985. This value is close to 2, which implies almost no autocorrelation.

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

   Answer: Based on the final model, the top 3 features contributing most towards the demand of shared bikes are the below:

   a. **atemp (coeff: 0.537)**: This feature is closely related to the temp variable. This makes sense as most users would like to use bikes as the temperature is higher.

   b. **hum (coeff: -0.284)**: This feature shows the humidity for each day. This feature negatively effects the number of users, as higher humidity usually deters users from riding bikes.

   c. **yr (coeff: 0.231)**: Denoting the year of operation, this feature as a positive coefficient on the bike demand, meaning that the number of bike users is increasing year to year.

General Subjective Questions

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

   Answer: Linear regression is a machine learning algorithm which falls under the supervised category and is used to predict numeric outcomes based on a linear equation containing one or more predictor variables and a target variable. The algorithm fits a linear equation using the Ordinary Least Squares and Gradient descent methods and generates the coefficients for the linear equation that are required to make predictions on the target variable.

   The generated linear equation for a dependant variable y using n predictor variables $X_1, X_2, \dots X_n$ will be in the below format. The equation shows how y changes with changes to one predictor variable while all other variables are constant.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_n X_n + \varepsilon$$

where,

- y is the dependent variable
- β_0 is the intercept or constant which denotes the value of target when all predictor variables are 0
- β_1, β_2, … β_n are the coefficients or slope of the predictor variables
- X_1, X_2, … , X_n are the predictor variables
- ε is the error

The linear regression algorithm contains 6 steps:

Step 1: Reading and understanding the data: Importing the data, describing the data, handling null values, standardizing the data, dropping irrelevant columns
Step 2: Visualizing the data: Exploratory Data Analysis (EDA), univariate, bivariate and multivariate analysis, pairplots, heatmaps, handling outliers
Step 3: Data preparation: Creation of dummy variables for categorical data, splitting data into training and test sets, rescaling the data, addition of constant to predictor dataset
Step 4: Training the model: Iterative training of model using a combination of Recursive Feature Elimination (RFE), statsmodels and scikitlearn(sklearn) libraries, checking for collinearity, statistical analysis of model
Step 5: Residual analysis: Calculation of predicted values, verification of model using plots to check normality of distribution of errors, check homoscedasticity of model using residual and predicted plot
Step 6: Making predictions using final model: Rescaling the test set, prediction of dependent variable using test set, evaluation of model by plotting actual values and predicted values of dependent variable in test dataset, derivation of final inferences from the model

For example the linear equation that is derived from the model is as below:

cnt = 0.231 x yr - 0.079 x holiday + 0.537 x atemp - 0.284 x hum - 0.178 x windspeed - 0.115 x spring + 0.049 x winter - 0.172 x light_rain - 0.086 x jul + 0.058 x sep

This means that there are 10 predictor variables that define the target variable. We will be able to predict the number of users by using the formula with the values of each predictor variable. The changes in the dependent variable with respect to any feature can be seen by changing the value of the variable while keeping others constant.

It is important to note the limitations of Linear Regression

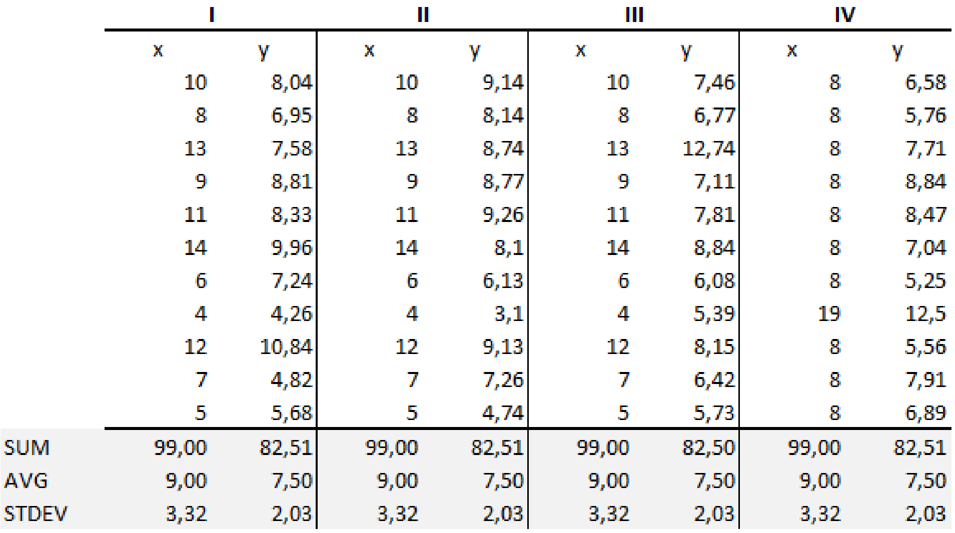
- Cannot predict outcomes when the relationship is not linear
- Can be used only for interpolation and not extrapolation
- Only shows correlation, does not guarantee causation

2. **Explain the Anscombe's quartet in detail.**

Answer: The concept of Anscombe's quartet was developed by the English statistician Francis Anscombe. The quartet consists of 4 datasets containing eleven pairs of predictor (x) and dependent (y) variables. These four datasets have the same sum, mean and variance for x and y. As the descriptive statistics are the same, the correlation for all 4 datasets is the same on paper. It is expected that a linear equation will predict the values of y well

However, when the datasets are plotted graphically, they tell a completely different story. The quartet demonstrates the importance of looking at the data visually using visualization techniques like graphs and not only using statistically.

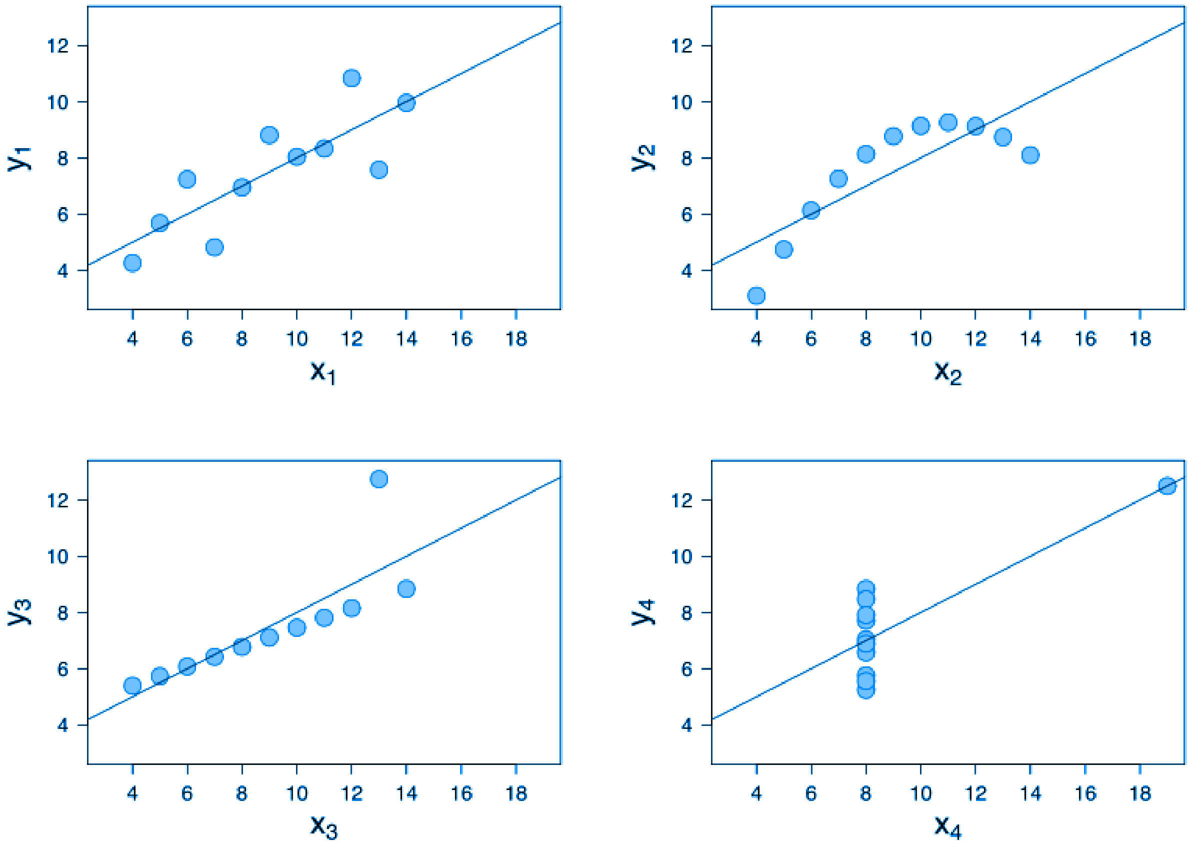
For example, consider the below 4 datasets.



	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

We can see that the sum, mean and standard deviation is the same for all datasets. Further the Rsquared for this data is also very high at 81.6%.

However, when we plot the above datasets on graphs, we can see the below:



Dataset I: The graph has a well-fitting linear model which defines the data well
Dataset II: The graph does not have a linear distribution
Dataset III: The graph shows a linear model which explains most points well but is skewed due to an outlier
Dataset IV: The graph shows data where we can see one far flung outlier is enough to create high correlation.

3. **What is Pearson's R?**

Answer: Pearson's R is a correlation coefficient which was developed by English mathematician Karl Pearson. This coefficient denotes the linear relationship between two variables. This coefficient can take the value 0 to 1. The formula to calculate this coefficient is as below

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

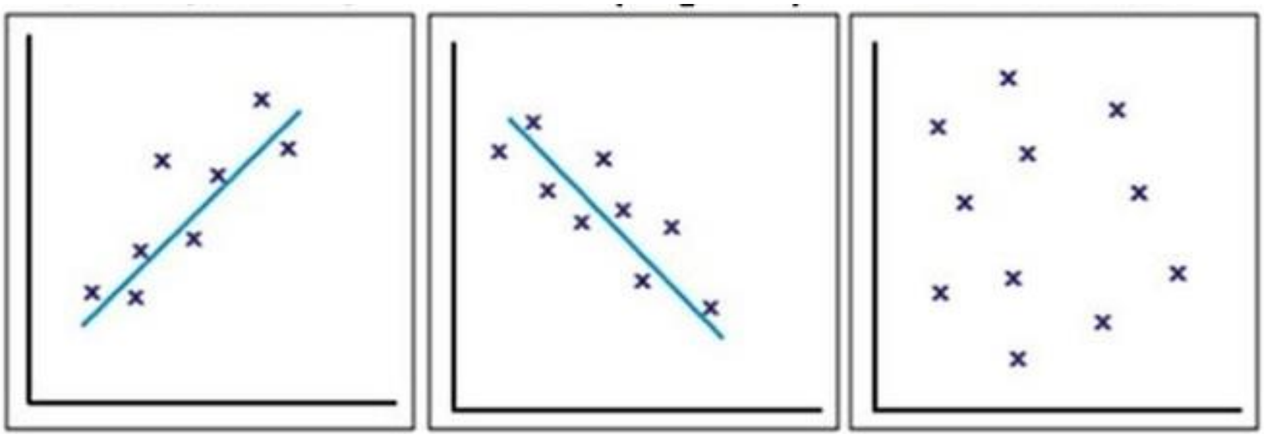
r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

When the value of correlation is positive, it denotes a directly proportional relationship, or a positive slope of the correlation line i.e. when one variable increases, the other variable increases as well

When the value of correlation is negative, it denotes an inversely proportional relationship, or a negative slope of the correlation line i.e. when one variable increase

When the value of correlation is high, or near 1, there is a strong linear relationship between the variables, and confidence of predictions is higher

When the value of correlation is low, or near 0, there is a weak linear relationship between the variables and confidence of predictions is lower



High positive correlation (r ~ 1) High negative correlation (r ~ -1) Low correlation (r ~ 0)

The validity of Pearson’s coefficient is based on the below assumptions:

a. Both variables are continuous
b. There is a linear relationship between both variables
c. There are no significant outliers in the data
d. Both variables are distributed normally

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

Answer: Scaling is a method which is used to standardize the value of the features which are to be used in regression. This is carried out at the data preparation stage of the regression to ensure all values in the dataset are at manageable levels of magnitude.

Scaling is done to ensure the below:

a. Features with high values do not have a higher effect on the final linear equation and are on a comparable scale
b. To improve the performance of the regression, as the model can easily arrive at the best fit using lesser steps during gradient descent

Scaling does not affect the model, but only scales the coefficients to match the corresponding scaling of the values.

Scaling can be done using normalization (Min-Max scaling) or standardization methods.

Normalization (min-max scaling) : the values are normalized to be between 0 and 1 using the below formula

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

where max(x) and min(x) are the maximum and the minimum values of the feature respectively

Standardization: the values are scaled to have zero mean and standard deviation. This is done using the below formula

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

where σ is the standard deviation of the feature vector, and x̄ is the average of the feature vector.

The difference between these approaches is as below

Normalization (min-max scaling)	**Standardization**
Scales values between 0 and 1	The bound of ranges is not certain
This method is highly affected by outliers	This method is less affected by outliers
This method is used when features are of different scales	This method is used to reduce the effect of high variance of data
The method uses minimum and maximum values in the feature for the scaling	This method uses mean and standard deviation to scale the data

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

Answer: The formula to calculate Variable Inflation Factor (VIF) is

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

An infinite value of VIF means that the value of R squared is 1. This happens when there is a perfect correlation between features. This happens when one feature is a direct linear derivative of another feature.

When this happens, we would need to remove the highly correlated variables one by one until VIF value is within acceptable limits

As a general rule of thumb, the value of VIF needs to be below the value of 5. A value of VIF higher than 5 means that the features are highly correlated.

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

Answer: The term Q-Q plot means a quantile-quantile plot, which is a probability plot which is used to check if two sets of data come from the same population. Such plots are very useful in the below cases:

- If two populations are of the same distribution

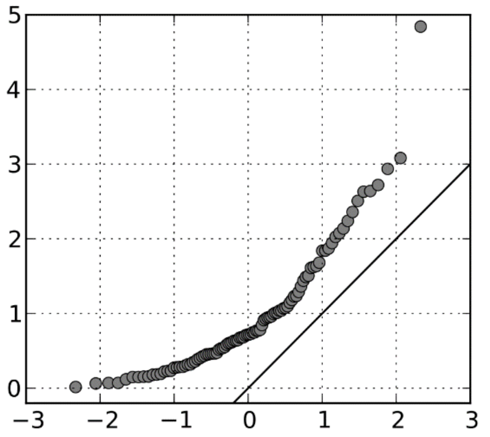
- If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it’s met using this.
- Skewness of distribution

This plot is commonly used in data science and quality control to verify assumptions and identify deviation from normal distribution.

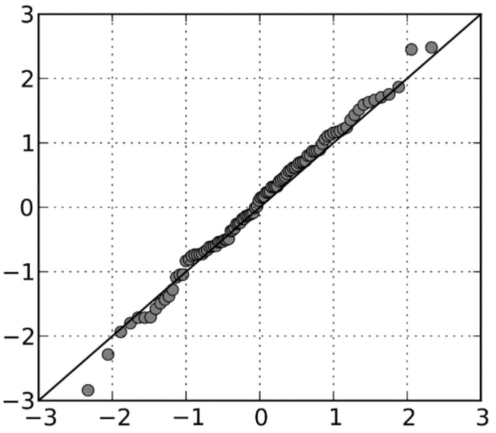
The graph is created by plotting the quantiles of the same datapoints from both datasets on a x-y plane. A quantile of a particular datapoint is the percentage of points which fall below the selected datapoint by value.

For example, when we say that the quantile of a particular datapoint is 70%, it means that the values of 70% of datapoints is lower than the selected datapoint.

In this graph, we also include a reference line with a slope of 1. Datapoints coming from the same population distribution will fall close to this reference line. If the plot lies farther from this reference line, the datasets come from different populations.



Q-Q plot showing nonlinear pattern meaning that the data is not normally distributed



Q-Q plot showing a strong linear pattern meaning that the data is normally distributed

Importance of Q-Q plots in linear regression

We can use Q-Q plots to check if the residuals from a linear regression model are normally distributed, which is a major assumption of such models. As we build our machine learning model, we need to ensure that we check the distribution of the error terms or prediction error using a Q-Q plot. If there is a significant deviation from the mean, we will then check the distribution of our feature variables and transform them to a normal shape.