Name: Kiran Prasad
Batch: EPGML C55 July 2023
Date: 10.10.2023

Assignment-based Subjective Questions

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

Answer:

Categorical variables have significant influence on dependent variable. 5 out of 7 variables used for predictions (after converting dummy variables to respective categorical variables) in the final model are categorical variables.

The business scenarios also supplement this dependency as bikes are normally used for short distance commuting and rather than wind speed or temperature on a specific day, the commuters generally decide on modes of transportation based on season/ chances of rain or snow/ month of year (again related to season) etc.

We can see that 'year' is a categorical variable with good influence on the dependent variable with a coefficient value of 0.23. This shows the demand for this mode of transportation is increasing year by year.

Final model statistics:

Variable category	Predictor variable	coef	P-value
Categorical	yr	0.2343	0
	holiday	-0.0919	0
	spring	-0.0716	0.001
	summer	0.0333	0.032
	winter	0.0887	0
	light_snow	-0.2929	0
	mist_cloudy	-0.0814	0
	December	-0.0445	0.012
	January	-0.0503	0.006
	July	-0.0504	0.007
	November	-0.0419	0.028
	September	0.0682	0
Numerical	temp	0.4377	0
	windspeed	-0.1586	0

Q-2:

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

Answer:

For a categorical variable having ‘m’ categories, only ‘m-1’ dummies are required to encode the ‘m’ categories. Hence we always drop one dummy variable using ‘drop_first = True’

For example: let ‘season’ is a categorical variable with categories as: Spring, summer, autumn and winter.

If we use 4 dummies for one hot encoding of this variable, it looks like as follows:

	Spring	**Summer**	**Autumn**	**Winter**
Spring	1	0	0	0
Summer	0	1	0	0
Autumn	0	0	1	0
Winter	0	0	0	1

Suppose we use only 3 variables instead of 4 – Spring, summer and autumn:

Spring : 100

Summer: 010

Autumn: 001

Winter: 000

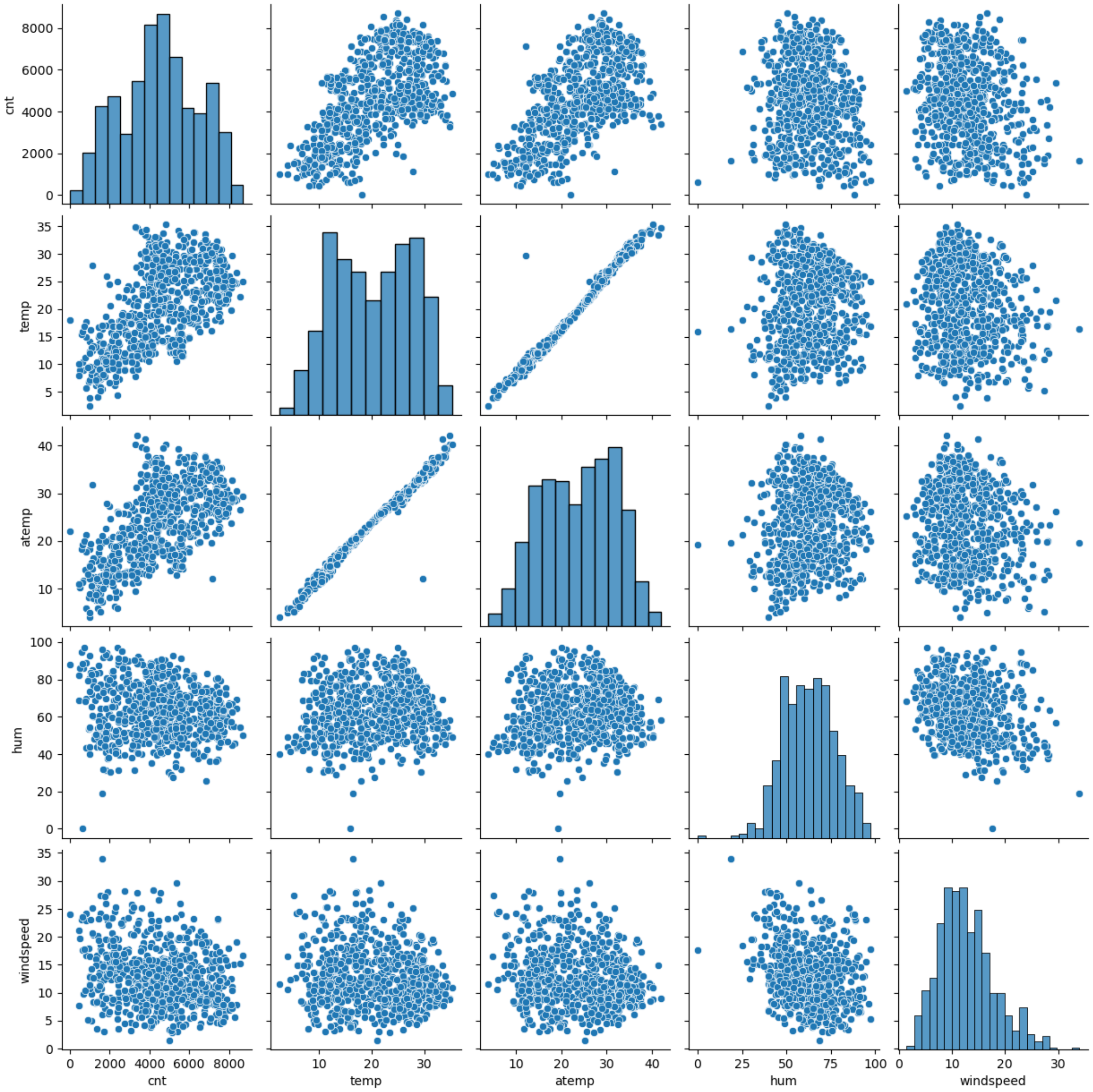
Even with 3 dummy variables, we are able to encode all 4 category levels. Thus we use only m-1 dummies to represent ‘m’ level categorical variable.

Q-3:

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

Answer:

Form the pairplot of numerical variables, it can be seen that ‘temp’ variable is having highest correlation with target variable.



Q-4:

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

Answer:

Validation of LR assumptions happens at two levels:

1. During model building:
    a. Linear correlation between target and predictor variables is visualized first to decide whether LR analysis is possible on given dataset or not.
    b. At each stage of model fine tuning by eliminating variables, the Variance Inflation Factor (VIF) is calculated to check multicollinearity among independent variables.
    c. A heatmap of independent variables in final model gives the autocorrelation among these variables.
2. During Model evaluation:
    a. Residual analysis focus on verifying the distribution of error terms – whether it is normally distributed with zero mean.
    b. A scatter plot of residual error Vs predicted target variable was used to check homoscedasticity of residuals.
    c. Durbin Watson test was used to check autocorrelation among residuals.

Q-5:

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

Answer:

Year, temperature and season are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

Variable category	**Predictor variable**	**coef**
Categorical	yr	0.2343
	holiday	-0.0919
	spring	-0.0716
	summer	0.0333
	winter	0.0887
	light_snow	-0.2929
	mist_cloudy	-0.0814
	December	-0.0445
	January	-0.0503
	July	-0.0504
	November	-0.0419
	September	0.0682
Numerical	temp	0.4377
	windspeed	-0.1586

General Subjective Questions

Q-1:

Explain the linear regression algorithm in detail.

Answer:

Linear regression is a statistical method for machine learning (ML) used for predictive analysis, which, as the name suggests, assumes a linear relationship between the target variable and the independent variables used for prediction.

Linear regression falls under the category of 'supervised learning' ML algorithms where the model is first trained using a labelled set of data. The model thus obtained is further used to predict the target value using unseen datasets.

Two types of linear regression:

1. Simple Linear Regression: The target variable is predicted using a single independent/predictor variable.
2. Multiple Linear Regression: The target variable is predicted using multiple independent variables.

Simple Linear Regression:

Basic type of regression where the relation between target and a single predictor variable is represented by a straight line.

Mathematical equation: $Y = \beta_0 + \beta_1X$

Where:

Y = target variable

β_0 = constant / intercept of line with y-axis

β_1 = coefficient of independent variable

X = independent/predictor variable

Aim of linear regression is to estimate the values for β_0 & β_1 such that the model 'fits' well with the data and predict the output with highest accuracy.

The best fit line (in other words, the values of β_0 & β_1) are found by minimizing the sum of squares of residual errors (RSS) .

RSS = sum[Y(actual) – Y(predicted)]^2

Strength of linear regression model is measured using a quantity called R^2 (coefficient of determination).

R^2 = 1 - (RSS / TSS)

Where, TSS = Total sum of squares = sum[Y(avg) – Y(actual)]^2

R^2 = 1 means all the predicted points lie on the straight line defined by the model

In other words, R^2 says how much variance in target variable is explained by the predictor variables used in linear regression model

Multiple Linear Regression:

An extension of simple linear regression were more than one independent variables are used for prediction of target variable. Here, we fit a 'hyper plane' of higher dimension, instead of a straight line. The dimension depends on the number of predictor variables used.

Mathematical equation: $Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots\ldots. + \beta_n X_n + \xi$

Increase in number of predictors will either keep the R^2 value same or it will increase. This doesn't mean the model has improved as there is chance for multicollinearity among the predictors which will lead to over fitting and model will fail while using new data sets. Hence 'Adjusted R2' parameter is used to measure the strength of MLR, which adds a penalty for increase in dimension.

Assumptions in Linear Regression:

1. There is a linear relationship between target variable and predictor variables
2. No multicollinearity among independent variables.
3. Errors/residuals are normally distributed
4. Errors have constant variance (homoscedasticity)
5. Errors have no autocorrelation

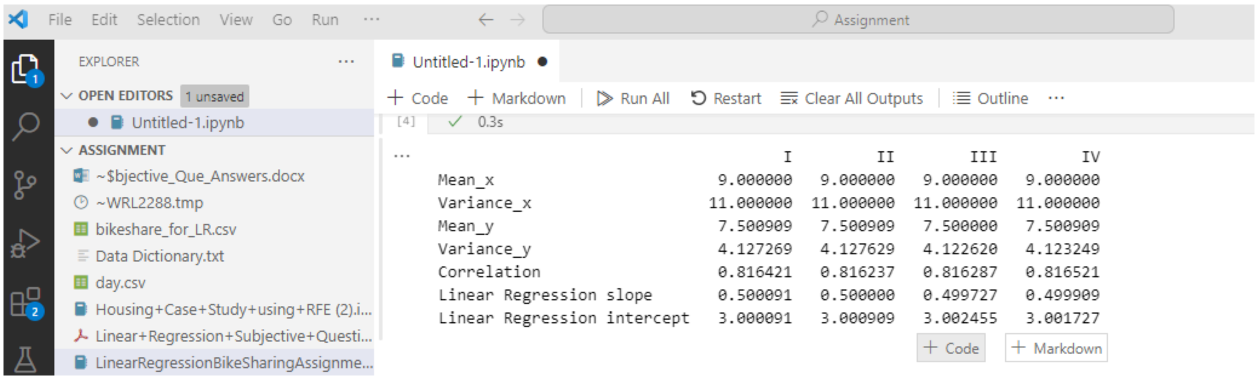
Q-2:

Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet is a set of 4 datasets created to illustrate the importance of visualization of data using EDA techniques rather than depending only on statistical parameters. These are 4 datasets having identical statistical measures like mean, variance, R squared value, correlation and also same linear regression line. But when we visualize the data using plots, we can see that each of the 4 datasets have entirely different plots.

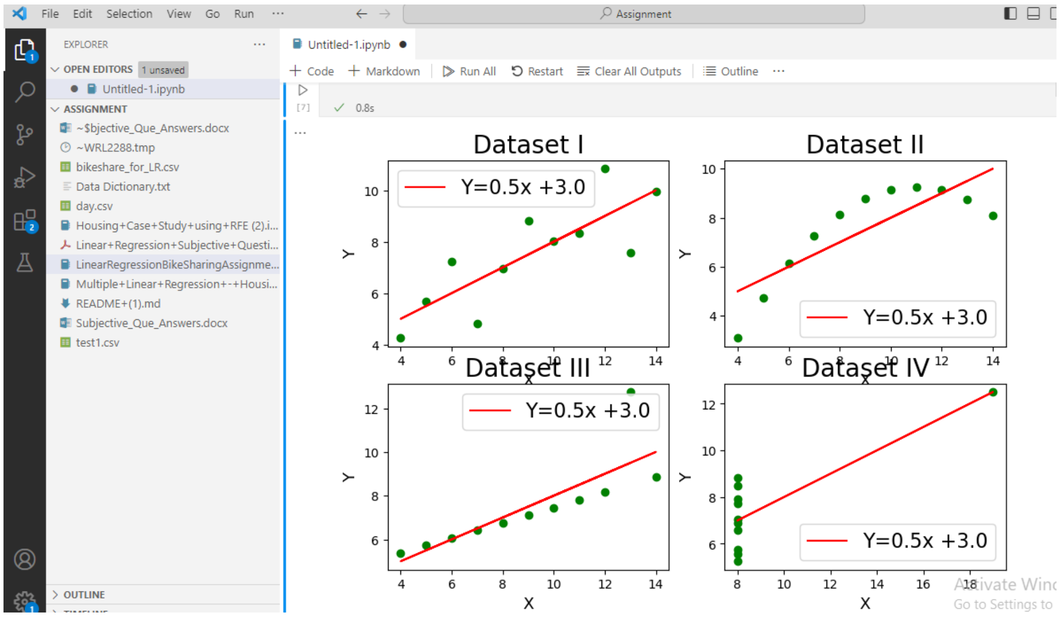
Statistical analysis of the Anscombe's quartet using python gives the following results:



```
                                   I         II        III        IV
Mean_x                      9.000000   9.000000   9.000000   9.000000
Variance_x                 11.000000  11.000000  11.000000  11.000000
Mean_y                      7.500909   7.500909   7.500000   7.500909
Variance_y                  4.127269   4.127629   4.122620   4.123249
Correlation                 0.816421   0.816237   0.816287   0.816521
Linear Regression slope     0.500091   0.500000   0.499727   0.499909
Linear Regression intercept 3.000091   3.000909   3.002455   3.001727
```

File Edit Selection View Go Run ···
Assignment
EXPLORER
OPEN EDITORS 1 unsaved
Untitled-1.ipynb
ASSIGNMENT
~$bjective_Que_Answers.docx
~WRL2288.tmp
bikeshare_for_LR.csv
Data Dictionary.txt
day.csv
Housing+Case+Study+using+RFE (2).i...
Linear+Regression+Subjective+Questi...
LinearRegressionBikeSharingAssignme...
+ Code + Markdown Run All Restart Clear All Outputs Outline
[4] 0.3s

And the plot of datasets is as follows:



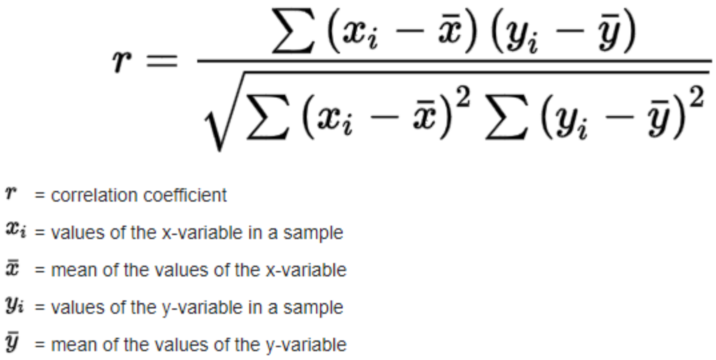
Q-3:

What is Pearson's R?

Answer:

Pearson's R or Pearson's Correlation Coefficient (PCC) is the measure of strength of linear relation between two variables. It calculates the effect of change in one variable w.r.t change in other variable.

Mathematical formula for Pearson's R is:



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

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

This is essentially a normalized measure of covariance and its value always lies between -1 and 1, calculated as ratio of covariance of two variables and the product of their standard deviations.

0<r<=1 → means positive correlation, both variables moves in same direction

-1<=r<0 → means negative correlation, both variables moves in opposite sides.

r = 0 → means no correlation

Q-4:

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

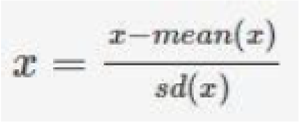
Answer:

Scaling is the process of converting the values of independent variables into standardized values such that all independent variables are having comparable values.

Scaling is performed in case of MLR where multiple independent variables are present with different scales thereby leading to building a model with very highly varying range of coefficient values, making it difficult to interpret the coefficients.

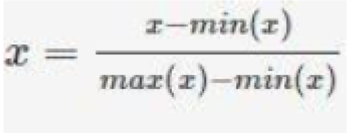
There are 2 types of scaling:

1. Standardization:
   Variables are scaled to have zero mean and SD =1.



$$x = \frac{x - mean(x)}{sd(x)}$$

2. Normalization / min-max scaling:
   Variables are scaled to have their values lying in between 0 & 1 using the min and max values available in data set.



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

Scaling can be performed without much implications as it will only affect the coefficients but not the statistical parameters of the model as a whole.

Q-5:

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

Answer:

Variation Inflation Factor (VIF) is the measure of multicollinearity of variables.

It says how well an independent variable can be represented by the rest of the independent variables used in a model.

High VIF means the selected variable is highly correlated with other independent variables.

OR
the selected variable can be represented as a LINEAR COMBINATION of other independent variable, making the selected variable a redundant one.

VIF = infinity means there is perfect linear correlation and the features of selected variable is 100% representable by the other variables.

Example:

For a blood donation camp, we have a dataset of students in a college which includes height, weight, BMI and quantity of blood that can be donated.

We know that BMI = weight in KG / (height in meters)^2

So if we try to build a MLR model with this data, on calculating VIF, we can see that VIF for BMI = infinity as BMI is completely representable with height and weight.

Q-6:

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

Answer:

Q-Q or Quantile-Quantile plots are used to compare a sample distribution against a standard theoretical distribution.

I.e. QQ plots are plots of quantiles of a sample distribution against quantiles of a theoretical distribution aiming to help in determining whether a sample dataset follows any particular type of probability distribution.

If datasets of comparison follows same distribution, then the plot will give a straight line with theoretical values in perfect straight line and the sample data overlapping the straight line with very less or no variance.

Q-Q plots are mainly used to determine the following details:

1. Whether two populations are following same distribution
2. Whether the residuals follow a normal distribution
3. Skewness of distribution