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?

Ans: We can conclude that the demand for bikes depends mainly on below variables: yr , holiday ,Spring, Light rain_Light snow_Thunderstorm,mist_cloudy, 3 ,5 ,8, 9, 10, 7. Demands increases in the month of 3, 5, 8 ,9, 10, 7. Demand decreases if it is holiday , Spring, Light rain_Light snow_Thunderstorm, Mist_cloudy

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

Ans: t is important in order to achieve k-1 dummy variables as it can be used to delete extra column while creating dummy variables. It is also used to reduce the collinearity between dummy variables .

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

Ans: Correlation between temp and atemp is higher which is 0.99

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

Ans: I have validated model on the basis of p-value. If p-value is high it means model is not working.

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

Ans: In the month of March, May, August , September, October and July demand increases

General Subjective Questions

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

Ans : A linear regression model describes the relation between a dependent variable and one or more independent variables. The dependent variable also called response variable and independent variables are also called predictor variables. Dependent variables denoted by Y and independent variables are denoted by X

M multiple linear regression model is as per below

yi=β0+β1Xi1+β2Xi2+…+βpXip+εi, i=1,⋯,n,

Where

- n is the number of observations
- yi is the ith response
- βk is the kth coefficient, where β0 is the constant term in the model.
- Xij is the ith observation on the jth predictor variable, j = 1, ..., p
- εi is the ith noise term, that is, random error

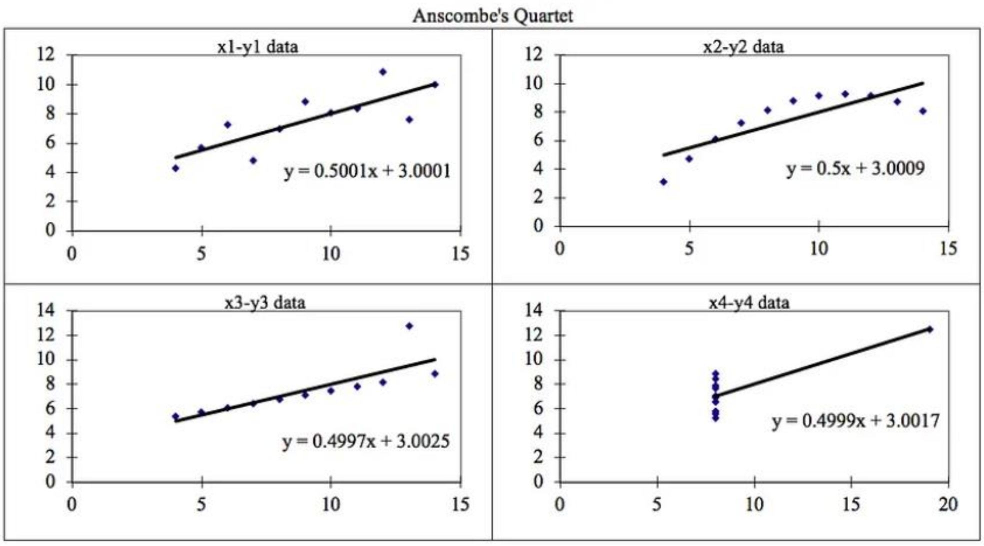
If a model includes only one predictor variable (p = 1), then the model is called a simple linear regression model.

In general, a linear regression model can be a model of the form

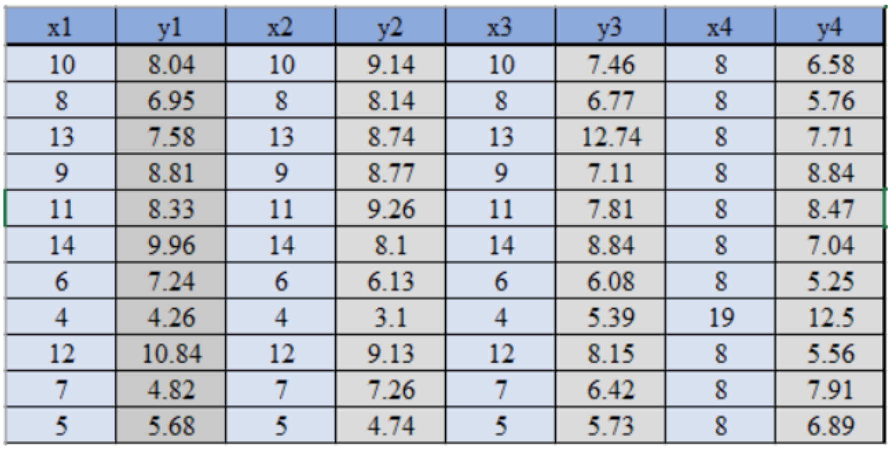
yi=*β*0+*K*∑*k*=1*βkfk*(*Xi*1,*Xi*2,⋯,*Xip*)+*εi*, *i*=1,⋯,*n*,

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

Ans : Anscombe's quartet were constructed by Statistician Francis Anscombe in 1973. it comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (*x*,*y*) points.



The basic thing to analyze about these data-sets is that they all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behavior irrespective of statistical analysis.



x1	y1	x2	y2	x3	y3	x4	y4
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

Four Data-sets

Apply the statistical formula on the above data-set,

Average Value of x = 9

Average Value of y = 7.50

Variance of x = 11

Variance of y =4.12

Correlation Coefficient = 0.816

Linear Regression Equation : y = 0.5 x + 3

However, the statistical analysis of these four data-sets are pretty much similar. But when we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represent the different behavior.

Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.

Data-set II — shows a curve shape but doesn’t show a linear relationship (might be quadratic?).

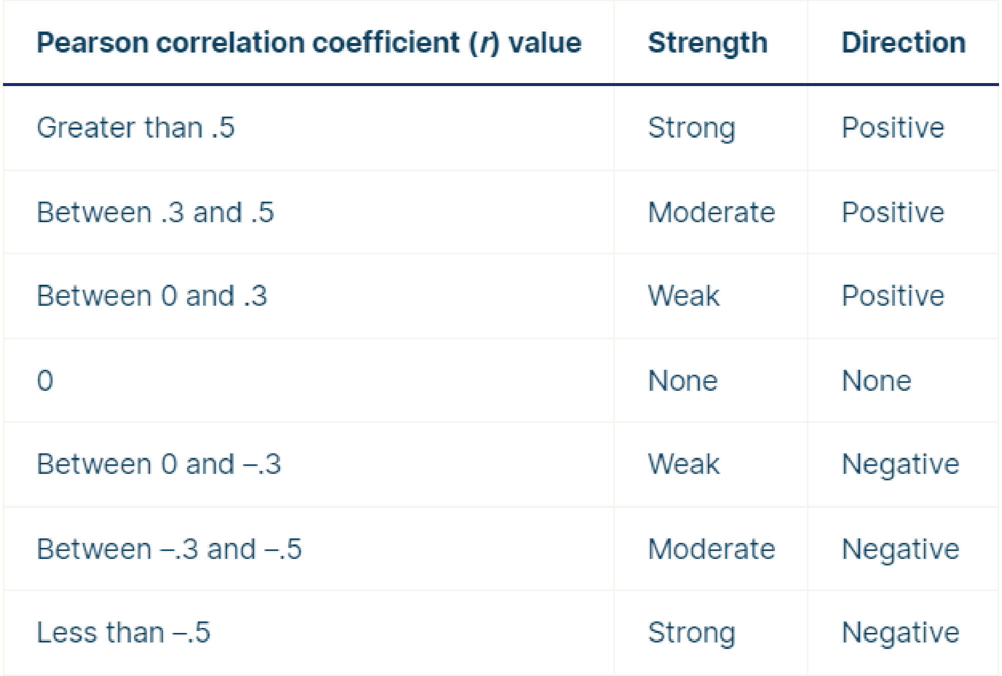
Data-set III — looks like a tight linear relationship between *x* and *y*, except for one large outlier.

Data-set IV — looks like the value of *x* remains constant, except for one outlier as well.

1. 3. What is Pearson’s R?

Ans; It is a Pearson correlation coefficient (r) is the most widely used correlation coefficient and is known by many other names which are Bivariate correlation, Pearson product-moment correlation coefficient (PPMCC), The correlation coefficient The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

Although interpretations of the relationship strength (also known as effect size) vary between disciplines, the table below gives general rules of thumb:



Pearson correlation coefficient (*r*) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

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

Ans : Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and 1. **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**).

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

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

Ans: VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables

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

Ans:Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.