Assignment-based Subjective Questions

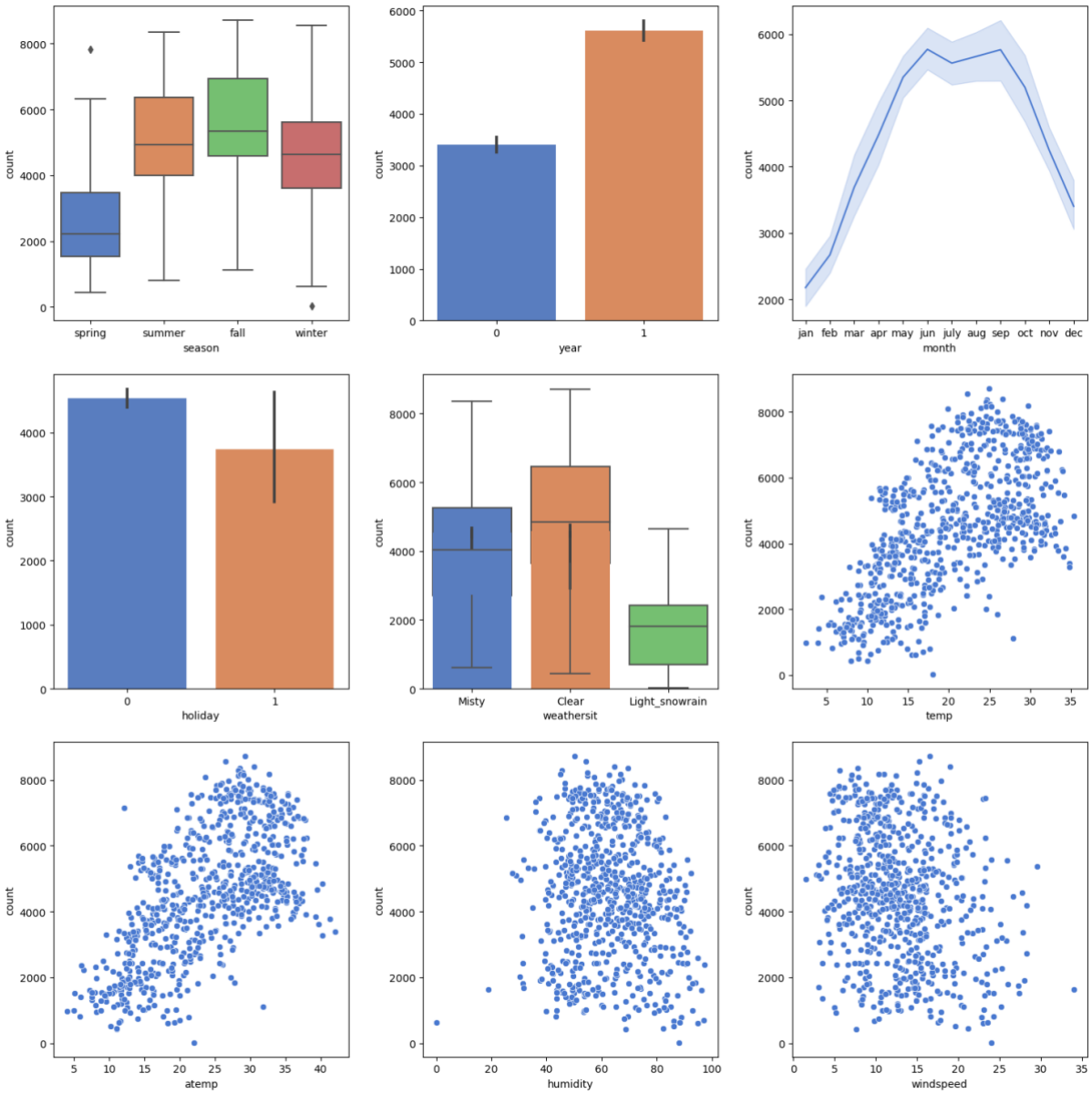
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

The following are the categorical variables:

i) season
ii) weathersit
iii) holiday
iv) month
v) year

The following inferences can be drawn:

- Bike sharing is the least in the **spring** season.
- The count variable is less during the **holidays**.
- The **fall** season is having the highest demand for rental bikes.
- Demand for rental bikes is increasing till the month of **June** staying there till mid **September**, then there is a fallback of demand.
- For the month of **September**, demand is highest for the year and then demand for rental bikes again decreases.
- Demand for bike rent is low in the beginning and end of the year. This may be due to bad weather conditions.
- Also, the number of rentals is more in 2019 than in 2018.
- The demand does not give a clear picture of whether it is a working day or a holiday.
- The demand increases when there is good **weathersit**.



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

drop_first = True reduces the extra variables created during the creation of the dummy variables. This implies it also reduces the correlation created between the dummy variables.

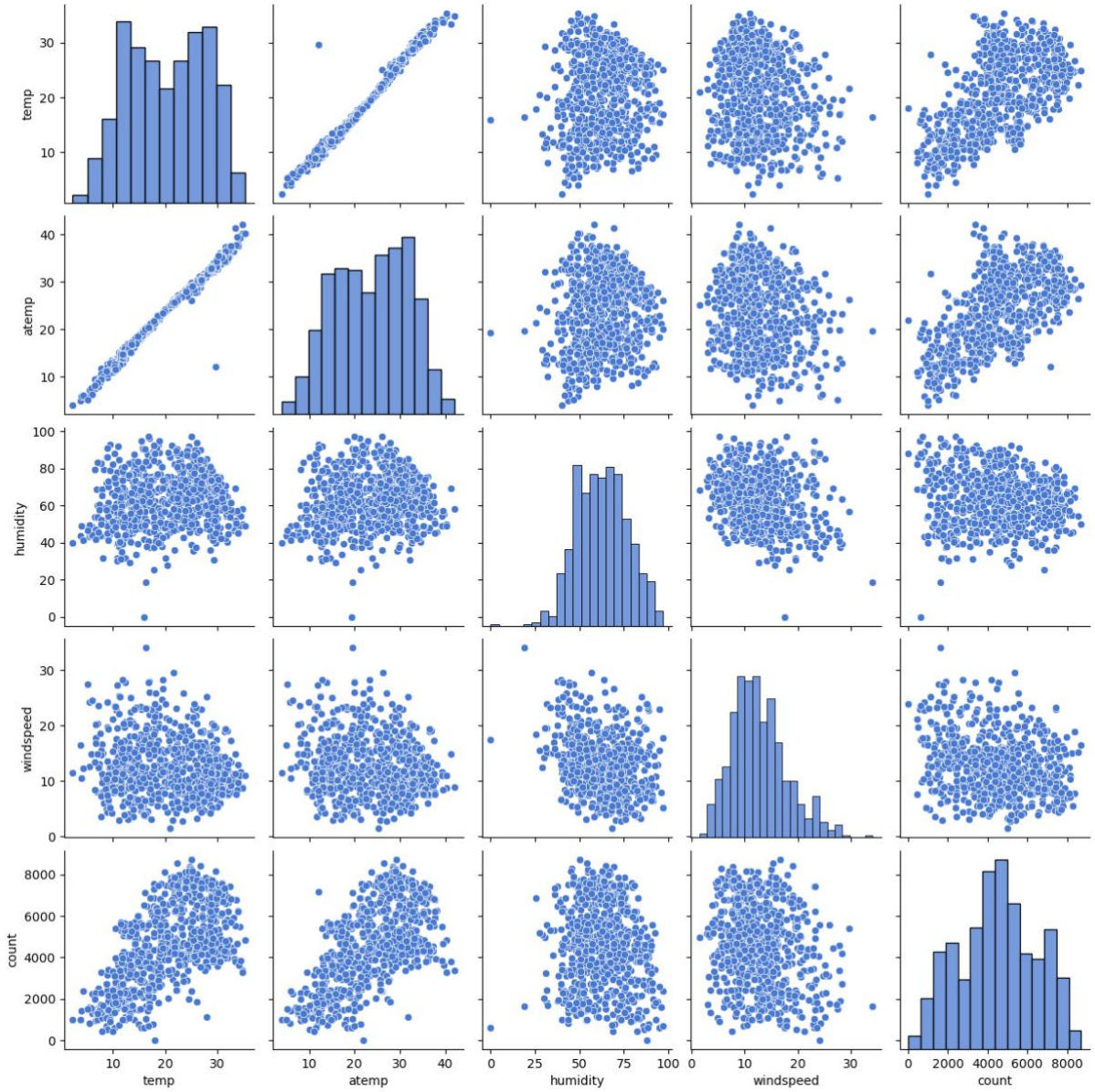
Syntax -

drop_first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not A and B, then it is obvious C. So we do not need 3rd variable to identify the C.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

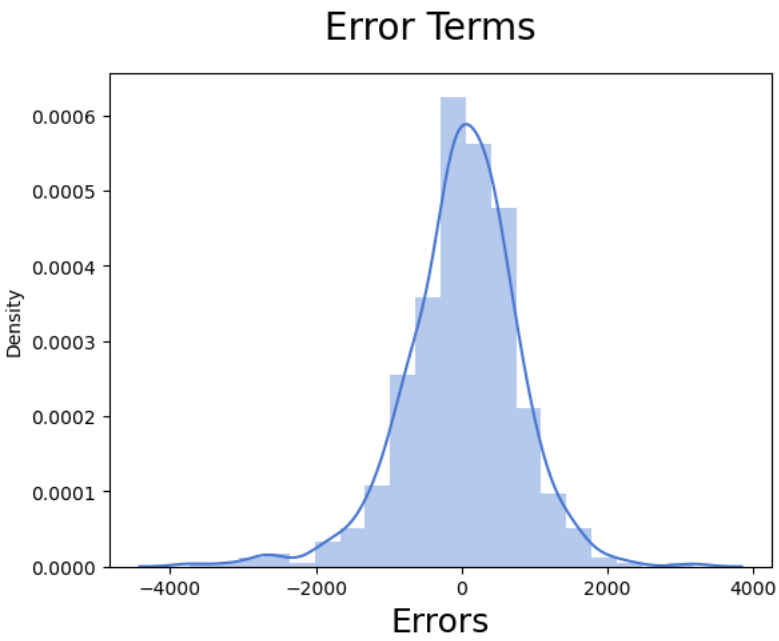
'temp' variable has the highest correlation with the target variable.



4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

To validate the assumptions, we should verify the residual to follow a normal distribution and mean = 0.

The below the chart verifies the normal distribution along with mean = 0:



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Top Predictor Variables:
Year (year):
Coefficient: 2034
Interpretation: A unit increase in the year variable increases the bike hire numbers by 2034 units.

Temperature (temp):
Coefficient: 3913
Interpretation: A unit increase in the temperature variable increases the bike hire numbers by 3913 units.

Light/Slow Rain (Light_snowrain):
Coefficient: -2535.0
Interpretation: A unit increase in Light_snowrain decreases the bike hire numbers by 2535.0 units.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
Linear Regression in Machine Learning is a statistical regression method which can be used to predict the analysis and visualize the relationship between the continuous variables. It is based on the equation:

$y = mx + c$
where, m = gradient
c is the intercept of the y-axis

Regression tries to find the best-fit line between the dependent and the predicted variables with minimal error.
It shows the linear relationship between the dependent variable (y-axis) and the independent variable (x-axis). It can be broadly divided into two types:

i. Simple Linear Regression – When the dependent variable is predicted using only one independent variable.
ii. Multiple Linear Regression – When the dependent variable is predicted using multiple independent variables.

Some use cases: We can use linear regression to predict the following:
• Predict the sales target for a company.
• Predict the scores of a student.
• Predict the change in stock price etc.

Assumptions -
The following are some assumptions about dataset that is made by Linear Regression model –
Multi-collinearity –
Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.
Auto-correlation –
Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.
Relationship between variables –
Linear regression model assumes that the relationship between response and feature variables must be linear.
Normality of error terms –
Error terms should be normally distributed
Homoscedasticity –
There should be no visible pattern in residual values.

2. Explain the Anscombe's quartet in detail. (3 marks)
Anscombe's Quartet is the model example to demonstrate the importance of data visualization to signify both the importance of plotting data before analyzing it with statistical properties. It comprises of four data-set and each data-set 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

Apply the statistical formula on the above data-set,

Average Value of x = 9

Average Value of y = 7.50

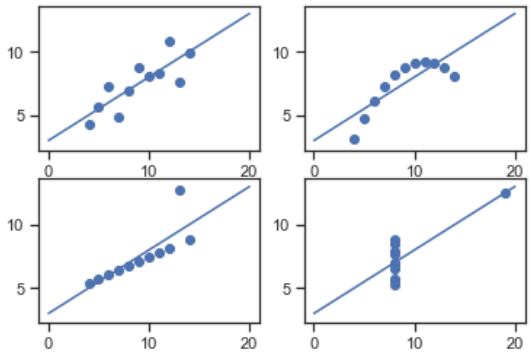
Variance of x = 11

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.

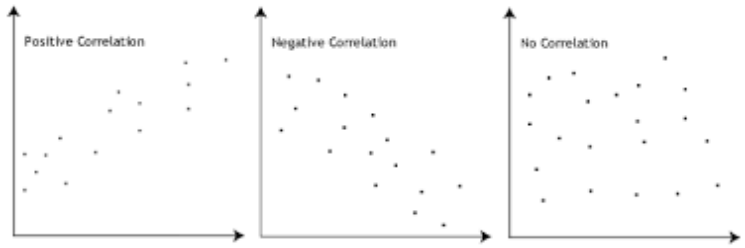


Graphical representation of Anscombe's quartet
Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.
Data-set II — shows a curve shape but doesn't show a linear relationship (might be quadratic?).
Data-set III — looks like a tight linear relationship 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.

3. What is Pearson's R? (3 marks)

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below (Picture taken from internet):



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is a method which is used to normalize or standardize some independent variables of the data set.

Scaling is performed at the pre-processing stage so that we can deal with the varying values in the entire dataset. Else if the units of the values are different and not standardized then it tends to give higher values for higher numbers and lower values for lower numbers.

Normalized scaling brings all the data in the range of 0 and 1. Minmaxscaler helps implement normalized scaling whereas standardised scaling replaces the values with z scores. One disadvantage of normalized scaling is it misses out on the outliers as it ranges from 0 to 1.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.