Assignment-based Subjective Questions

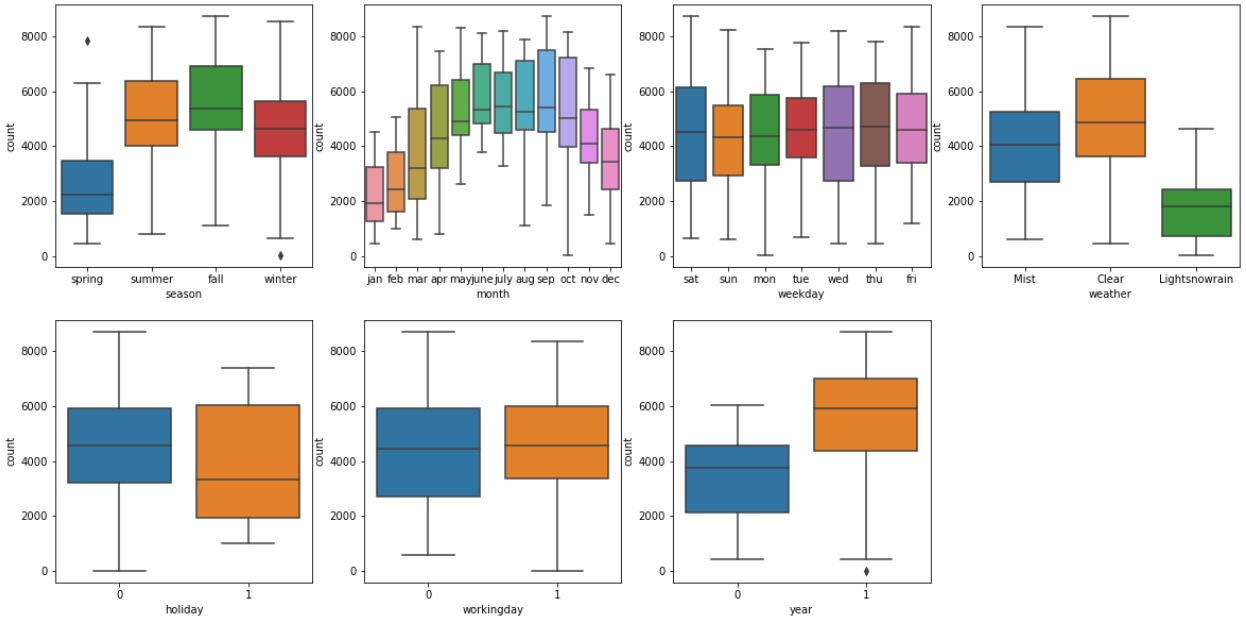
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer:

The available categorical variables in dataset are Season, month, holiday, weekday, year, working day and weather.

Effect of categorical variable on dependent variable:

a) The year 2019 having increased booking count than 2018, means the business increasing on a yearly basis.
b) The season Fall having higher booking count and summer season is second highest season in no. of booking.
c) Clear weather is one of the convenient factor for bike ride. As evidence, The clear weather having highest booking count in given dataset
d) Holiday having lesser no. of booking than non-holiday. We can assume, holiday used for taking rest and to spend time with family.
e) Working day and non – working day is not showing much difference.
f) All weekdays having moderately same no. of booking, Only on Sunday we can see lesser count than other days.
g) The month Jan, Feb, Mar, Nov and Dec is having lesser number of booking count.



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

Answer:

Get_dummies function used to encode (One Hot Encoder Method) the categorical variable into a binary vector representation for the understanding of machine learning model.

Drop_first = True will removes the first column which is created for the first unique value of a column.

drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Even if we drop the extra column, the information provided by that categorical column will not change.

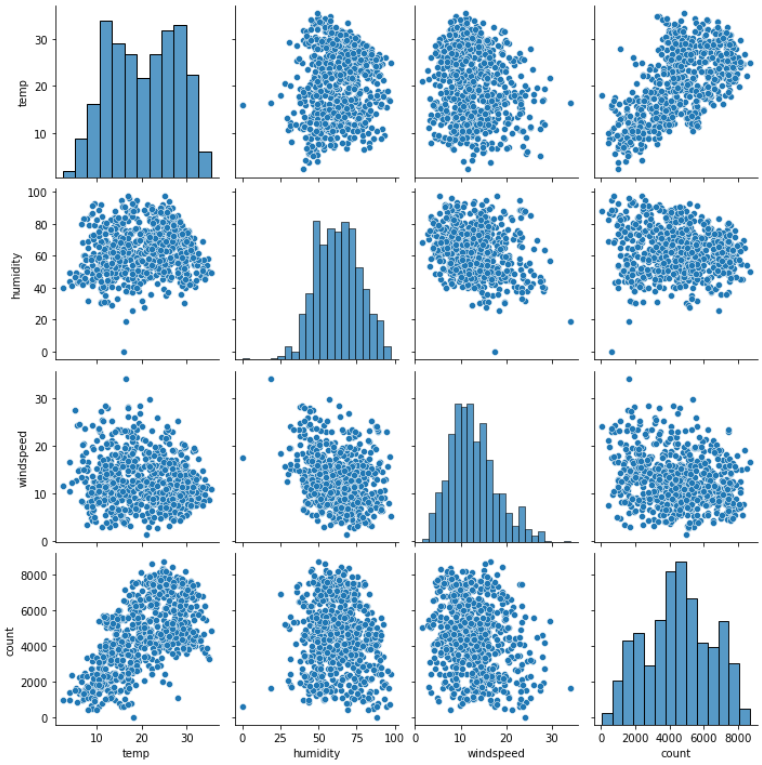
For p catergorical variable, p-1 dummy is being created.

Syntax - pd.get_dummies(df.<categorical_column>,drop_first=True)

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

Answer:

The temp (temperature) numeric column is having highest positive correlation with target variable.
Correlation between temp and cnt(target variable) = 0.63

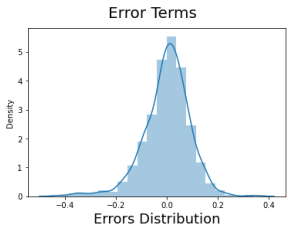


4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

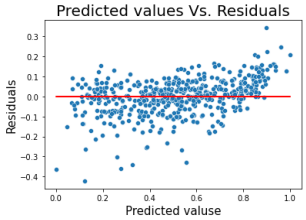
Answer :

Linear Regression Assumption :

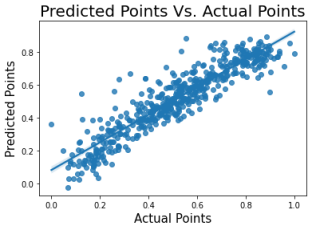
1) Linear relationship between target variable and the predicted variable.
2) Error term(difference between actual and predicted variable) are normaly distributed with mean equal to Zero.



3) Error terms are independent of each other means no visible pattern between residuals and predicted values.



4) Error term have constant variance(homoscedastic).



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

Answer:

Top 3 feature contributing significantly:

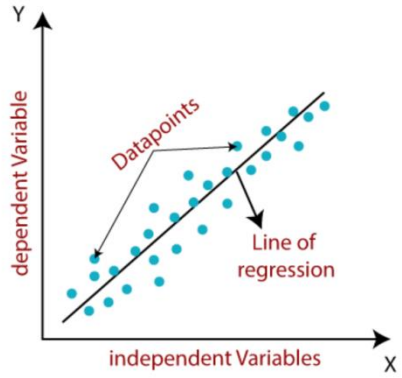
1) Temp
2) Winter
3) September

General Subjective Questions :

1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression makes predictions of continuous numeric variables such as sales, salary, age, product price, etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression.



Y = B0 + B1(x) + E,

Y – Target variable, B0 – Intercept of the line, B1 – Scalar factor for each input value, E – Random Error.

There are two types of linear regression:

1) Single Linear regression – when number of independent variable is one.

2) Multiple Linear regression – When number of independent variable is greater than one.

A regression line can show a two type of relation ship:

1) Positive Linear Relationship : The increase in X – independent variable the Y – dependent variable also increases.
2) Negative Linear Relationship : The Increase in X – Independent variable, the Y – dependent variable will decreases.

The best fit line means the error between predicted values and actual values should be minimized. The best fit line will have the least error.

Assumption of Linear Regression model:

1) Linear relationship between target variable and the predicted variable.
2) Error term(difference between actual and predicted variable) are normaly distributed with mean equal to Zero.

3) Error terms are independent of each other means no visible pattern between residuals and predicted values.
4) Error term have constant variance(homoscedastic).

The Best model can be found with the help of below parameters:

1) R squared values - It measures the strength of the relationship between the dependent and independent variables on a scale of 0-100%.
2) Probability of F-stats – The lesser the value , the model will be significant.
3) P-Value - <0.05 is a significant value for optimal model.
4) VIF(Variance inflation factor) – Says linear relationship between independent variables.

2. Explain the Anscombe's quartet in detail. (3 marks)

Answer:

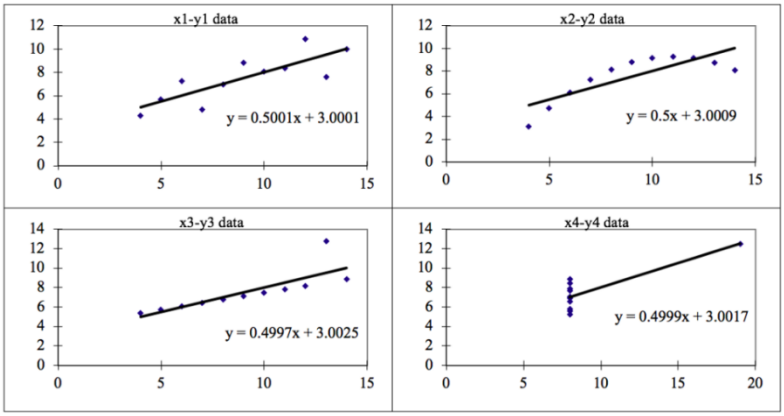
Anscombe's Quartet can be defined as a **group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built**.

They have very different distributions and appear differently when plotted on scatter plots.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

When these models are plotted on a scatter plot,



The four datasets can be described as:

Dataset 1: this fits the linear regression model pretty well.

Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.

Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model

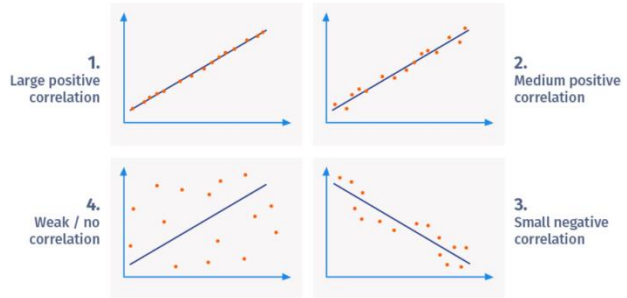
Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

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

Answer:

Pearson correlation coefficient or Pearson's correlation coefficient or Pearson's r is defined in statistics as the **measurement of the strength of the relationship between two variables and their association with each other**.

Pearson's correlation coefficient calculates the effect of change in one variable when the other variable changes.



It looks at the relationship between two variables. It seeks to draw a line through the data of two variables to show their relationship. The relationship of the variables is measured with the help Pearson correlation coefficient calculator. This linear relationship can be positive or negative.

For example:

Positive linear relationship: In most cases, universally, the income of a person increases as his/her age increases.

Negative linear relationship: If the vehicle increases its speed, the time taken to travel decreases, and vice versa.

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

It 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.

Scaling Advantage – Better interpretability and for faster calculation at the backend

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.

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).

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

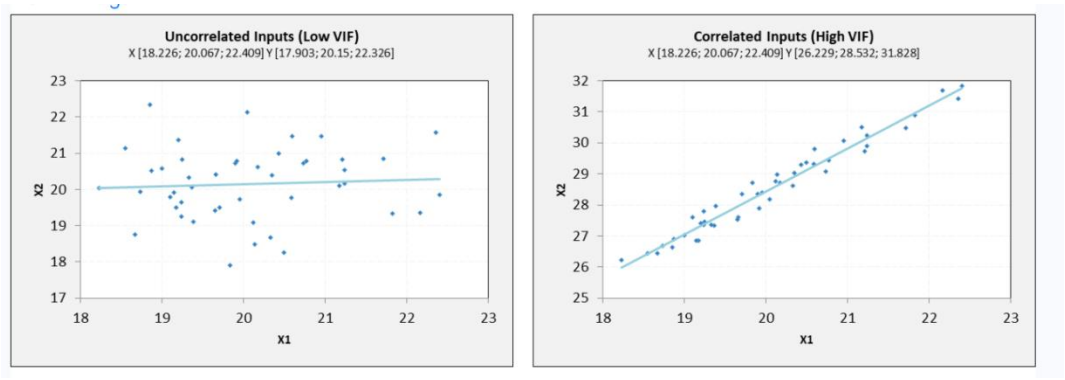
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer:

VIF value tells about the correlation between all the independent variable available in the dataset

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).



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

Answer:

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile.

Uses:

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.

Importance:

Probability distributions are essential in data analysis and decision-making. Some machine learning models work best under some distribution assumptions. Knowing which distribution we are working with can help us select the best model. Hence understanding the type of distribution of feature variables is key to building robust machine learning algorithms.