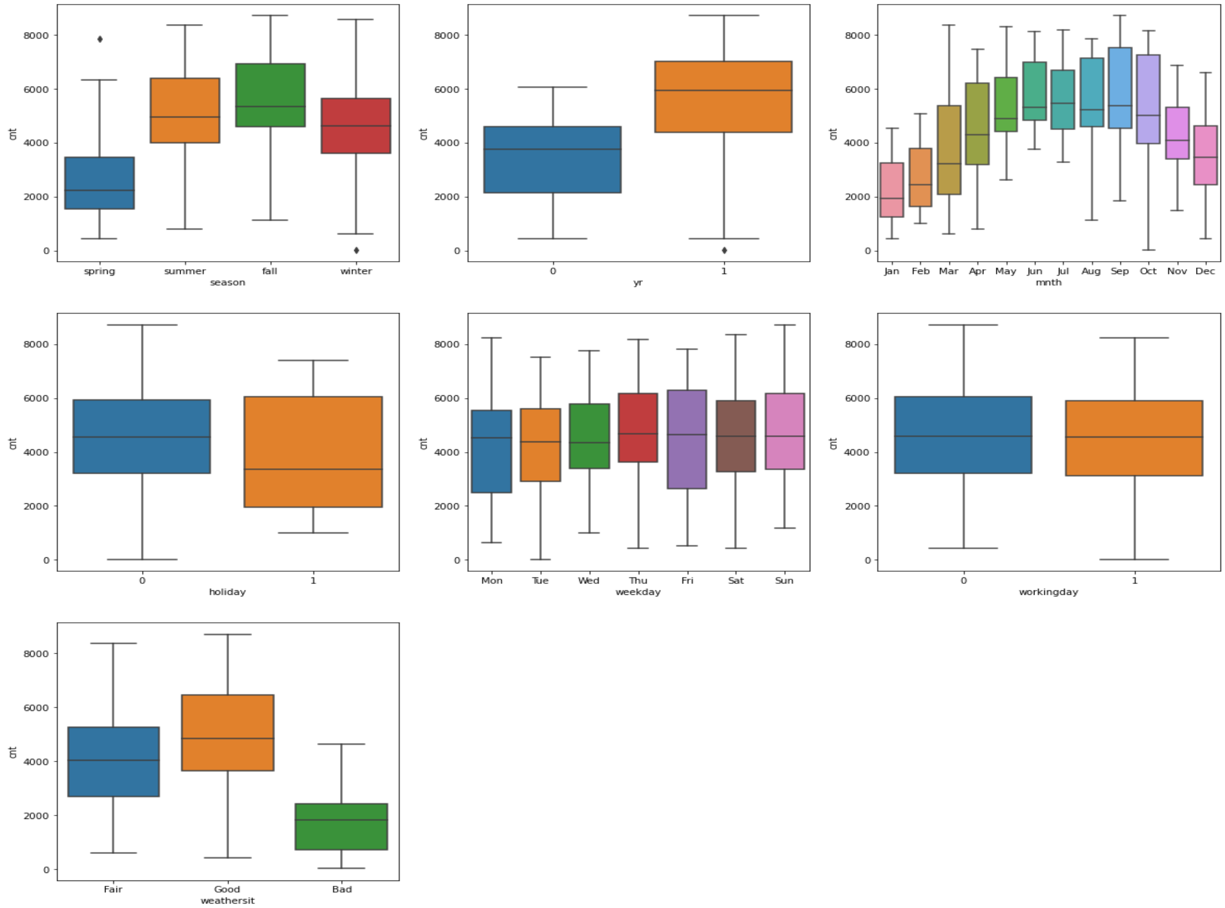
Assignment-based Subjective Questions

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

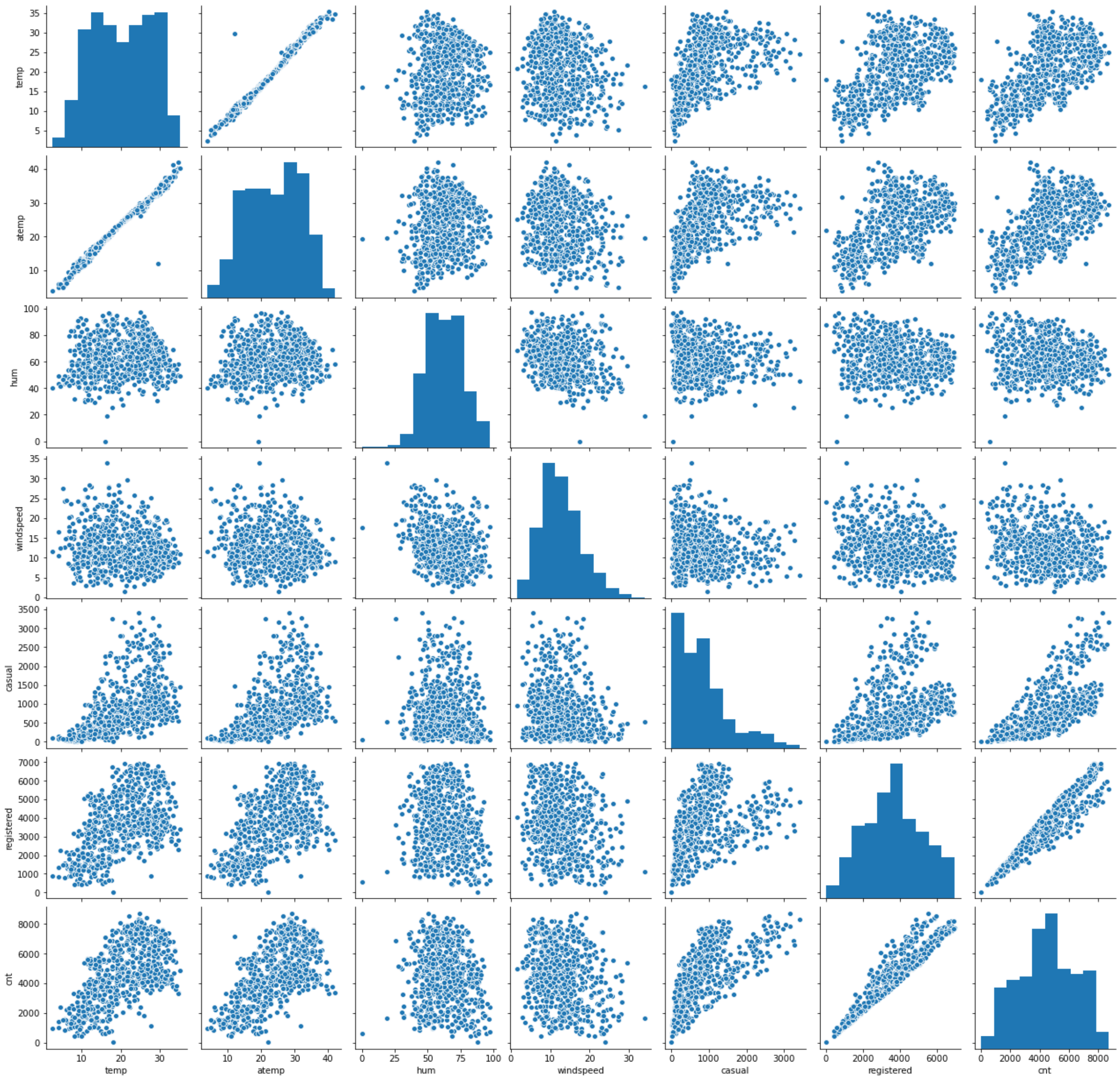
- We can see that count of renting bikes are more in year 2019 than 2018
- Count of renting bikes is less in spring season, whereas summer and fall season has higher count of renting bikes.
- Month of September, October and August have higher booking count, whereas January, December and November has lower count of renting bikes
- We can see that most renting bike bookings happened on working days rather than holidays
- We can see that most bookings for rental bikes are happened from Thursday to Monday and slightly low on Tuesday and Wednesday
- We can also see that the bookings for rental bikes are happening more when the weathersit is good or fair

Please find below the plot with cnt for categorical variables



2. Why is it important to use drop_first=True during dummy variable creation?
Answer:
It is important to use as it reduce the Extra column while creating the dummy variable. For example, let's say if we have column having "male" and "female" categories. If we create dummies we should get 2 dummies one is "male" another one is "female". However, we don't need two dummy columns to represent the same data if 1 is male obviously 0 is female. So it is better to use **drop_first=True** since it reduces an extra column. If we have to create n dummies for a given column, the number of dummies we get is n-1. It also reduces the correlation between the dummy variables.

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



From the above plot, we can see that with target variable ‘cnt’ highest correlation is with ‘registered’ column, however, since casual + registered = ‘cnt’. We shouldn’t be considering this variable with highest correlation. So, the next highest correlation is with ‘temp’ and ‘atemp’ both having similar linear patterns with ‘cnt’.

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

Assumption 1: The Dependent variable and Independent variable must have a linear relationship.

To validate this assumption we should check pair plot to see linear patterns in the data and see whether a linear relationship between dependent and independent variables.

Assumption 2: Residuals must be normally distributed. If it is violated, it causes problems with calculating confidence intervals and various significance tests for coefficients.
We can use distplot for residuals and see if it is normally distributed.

It can be fixed by outlier treatment and nonlinear transformation such as log

Assumption 3: Error terms are independent of each other

We can plot a scatter plot between X and residual in case of 1 independent variable. In case of multilinear regression, we can plot with an index having the same length of data frame

Assumption 4: Multicollinearity, all independent variables are independent of each other and have no correlation amongst each other.

If it is violated, coefficient estimates will be unstable and can cause the coefficients to switch signs.

We can removing variables with VIF (Variance inflation factor) larger than 10 or by PCA (Principal component analysis)

Assumption 5: Error term should have equal variance across different observations. (Heteroscedasticity)

In the case of heteroscedasticity is violated, it leads to wrong standard errors of the coefficients, which results in wrong t-statistic and p-values

It can be identified if residual vs fitted values has specific pattern, it means has non-constant variance

Assumption 6: Autocorrelation
If errors are correlated and not independent, it’s said to have auto-correlation. In other words, error of t instance bears an impact from t−1

It can be tested by durbin-watson test. If it is between 1.5 - 2.5, it’s good. Less than 1.5 is negative autocorrelation, more than 2.5 is positive autocorrelation. Although 2 is considered as ideal value.

It can be fixed by adding a lag variable and seasonality related dummy variables.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Answer:
The independent variables ‘atemp’, ‘yr’ and ‘windspeed’ are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Question

1.Explain the linear regression algorithm in detail.

Answer:

Linear regression defined as the model which analyzes the linear relationship between a dependent variable with a set of independent variables. Here, y is dependent variable and X1, X2, X3…are independent variables.

If the number of independent variable is 1 it is called simple linear regression. In order to perform simple linear regression for a data, we need to make few assumptions:

1. Linear relationship between independent and dependent variable
2. Error terms are normally distributed with mean zero
3. Error terms are independent of each other
4. Error terms have constant variance (homoscedasticity)
5. Multicollinearity
6. Autocorrelation

Once we done with the assumptions, in a simple regression problem (a single x and a single y), the form of the model would be:

y = B0 + B1*x

In higher dimensions when we have more than one input (x). The representation therefore is the form of the equation and the specific values used for the coefficients (e.g. B0 and B1 in the above example).

We will then split the data to train and test (unseen data) datasets. We will use the train data set to train the model to make predictions.

We have mainly 2 methods to train our model

1. Ordinary least square method
2. Gradient descent method

In ordinary least square method, the main motto is to minimize the sum of the squared residuals.

We can calculate residuals by squaring the difference between actual and predicted value.

This approach treats the data as a matrix and uses linear algebra operations to estimate the optimal values for the coefficients. It means that all of the data must be available and you must have enough memory to fit the data and perform matrix operations. This procedure is very fast to calculate.

In gradient descent, when there are one or more inputs you can use a process of optimizing the values of the coefficients by iteratively minimizing the error of the model on your training data.

This operation is called Gradient Descent and works by starting with random values for each coefficient. The sum of the squared errors are calculated for each pair of input and output values. A learning rate is used as a scale factor and the coefficients are updated in the direction towards minimizing the error. The process is repeated until a minimum sum squared error is achieved or no further improvement is possible.

Making predictions with linear regression:

We can use Statsmodel.api and sklearn libraries to build our linear model.

Step 1: Reading and understanding the data

We should read and see the shape, size and datatypes of each column in the dataset. Also, we can check if there are any null values present in the dataset

Step 2: Modifying the variables

We should identify if there any errors in datatypes and categorical variables in the data and try to modify them.

Step 3: Visualizing the data

We should visualize the data and understand how the independent variables vary with dependent variable. Also, we can see the pair plot to understand which independent variables have linear relationship with dependent variable

Step 4: Preparing the data

We should prepare the data by creating dummies for the categorical variables and dropping irrelevant columns.

Step 5: Performing linear regression

We can perform this using Stats model or by Sklearn library. We should select our training data and scale if required and fit to the model. By sklearn library, we will get the model and by statsmodel we will get the detailed stats related to the coefficients of out model such as p-value, F-statistic, R Squared, Adj. R Squared, etc. which will be helpful in determining the significance of each coefficient. We need to check VIF (Variance inflation factor) to determine the correlation between independent variables. We should drop the variables having >10 value for VIF and p-value >0.05. We should not drop multiple variables at a time in case if 2 variables have more than 0.05 p value and >10 VIF. We should drop one by one as there is a chance another variable VIF or p value decreases in case if we drop one variable. We perform this task again and again until we get the significant parameters.

Step 6: Residual Analysis

Once we build our model, now it’s time to check if our assumptions are correct.

We will plot a distplot for residuals (y actual- y predicted) to check if it is normally distributed. Also, we can check for Autocorrelation using durbin Watson test. Also, we need to check if error terms have constant variance (homoscedasticity)

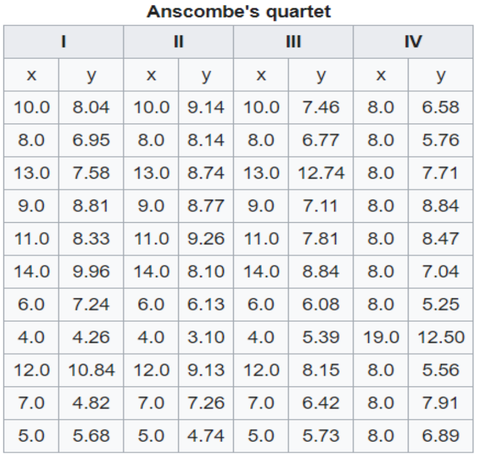
Step 7: Model Evaluation

We can plot y_test and y_pred to understand the spread and how well our model is predicting the output with test data.

Finally, we will achieve at a model which is able to explain the variance in the data based on the r2 score and we can also check which describes the target or independent variable.

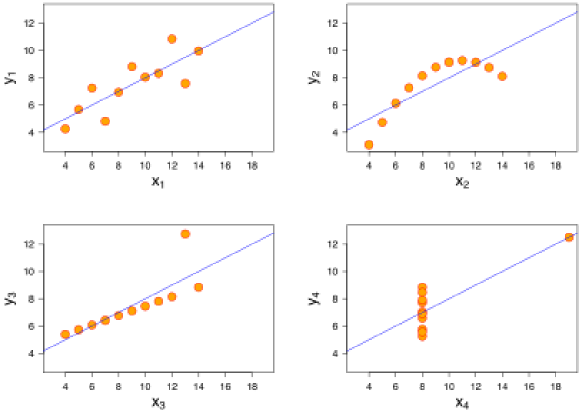
2. Explain the Anscombe’s quartet in detail.
Answer:
Anscombe’s quartet consists of 4 datasets that have nearly identical simple statistical properties, however, they have different distributions and appear very different when graphed.

It determines the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.



Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



From the data above, all four of these data sets have the same variance in x, variance in y, mean of x, mean of y, similar correlation and linear regression. However, the plots vary from each other.

Plot 1 (Top left): This fits the linear regression model pretty well

Plot 2 (Top right): This couldn’t fit the linear regression model on the data quite well as the data is non-linear

Plot 3 (Bottom left): In the third graph, the distribution is linear, however, the regression line is offset due to one outlier present in the data

Plot 4 (Bottom right): In the fourth graph shows an example one highest point can form a linear relationship even though the other data points do not indicate any relationship between the variables.

Finally, we can say that it is important to visualize the data before proceeding to analyze the data and model building.

3. What is Pearson's R?
Answer:

Pearson's R is a numerical summary of the strength of the linear association between the variables.

The Pearson's correlation coefficient varies between -1 and +1 where:

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)

r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

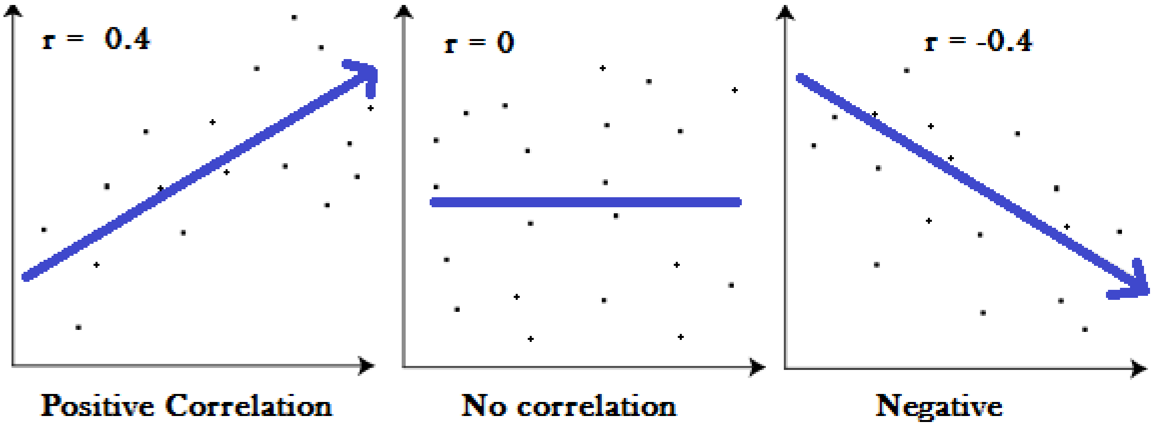
r = 0 means there is no linear association

A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.

A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed.

Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

Please find the graph below to under the Pearson's R



Pearson's correlation coefficient formula is given by

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

Sample correlation coefficient

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Sx and sy are the sample standard deviations, and sxy is the sample covariance.

Population correlation coefficient

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

The population correlation coefficient uses σx and σy as the population standard deviations, and σxy as the population covariance.

In simple terms, it answers the question, Can I draw a line graph to represent the data?

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

Scaling is a step of data pre-processing which is applied to independent variables to normalize the data within a particular range. It helps in speeding up the calculations in the algorithm.

For the collected data, if there is vast difference in range say few ranging in thousands and few ranging in tens, it makes an assumption that higher ranging numbers have superiority and thus it makes a priority to this variables. Machine learning algorithms works on numbers and it doesn't know the units of the variables represent. Thus, features scaling is important to bring every feature to the same scale and it also improves the calculation speed in algorithm.

There are several type of scaling techniques available, the mostly used techniques are Normalized scaling and standardized scaling.

Normalized scaling: It is also known as Min-Max scaling. It brings all the values between 0 and 1.

Standardized scaling: Standardization replaces the values with Z-scores. It brings all the data into standard normal distribution which has mean 0 with standard deviation 1

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

Answer:

Variance inflation factor helps in measuring the multicollinearity among the independent variables.

$$\mathbf{VIF = 1/Tolerance = 1/1\text{-}R_i^2}$$

R_i^2 represents the correlation coefficient of i^{th} independent variable with remaining independent variables.

If $R_i^2 = 1$, then the VIF = infinity, it means that the independent variable is highly correlated with all the other independent variables and Correlation coefficient = 1.

If $R_i^2 = 0$, then the VIF = 1, it means that the independent variable has no correlation with all the other independent variables and Correlation coefficient = 0.

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

Answer:

Q-Q plot is known as Quantile – Quantile plot. It is a graphical tool which helps us understanding which type of distribution the data came from theoretical such as, normal, uniform or exponential distribution. It also determines if two data sets come from populations with a common distribution.

In order to create this plot, we need to get the quantiles of the data (Actual quantiles) and we will plot these quantiles on the distributions (Normal, uniform, exponential, etc) and we should get the Z value corresponding to these quantiles which will give us theoretical data. Once we get the theoretical and actual values, we can plot a scatter plot with Actual vs theoretical values of the data. If we are able to fit a straight line, then our assumption of the distribution is correct, if not we can check for another distribution.

This helps in a scenario of linear regression when we have training and test data sets separately, then we can confirm it by using Q-Q plot that both data sets are from populations with a common distribution.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set

If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

Y-values < X-values: If y-quantiles are lower than the x-quantiles.

X-values < Y-values: If x-quantiles are lower than the y-quantiles.