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: There are in total 7 categorical variables and below is their effect on the dependent variable:

1. Season - Most of the bikes were booked in the summer and fall season whereas the spring season has very few numbers of bookings.
2. Year – The bookings of bike had increased from 2018 to 2019 which shows that the demand of bikes increases year on year.
3. Month - Most of the bikes were booked in the month of September. Also is the observed that mid-year has maximum number of bookings whereas beginning and end of year has minimum bookings.
4. Holiday – There is a fall in the number of bookings if it is a holiday.
5. Weekday - All the days of the week had almost equal number of bookings and thus weekday has no effort of the dependent variable.
6. Workingday – There are almost equal number of bookings in working as well as non-working days, thus workingday has no effort of the dependent variable.
7. Weathersit - Maximum number of bookings are done when the weather was either very good or normal and the sky is clear.

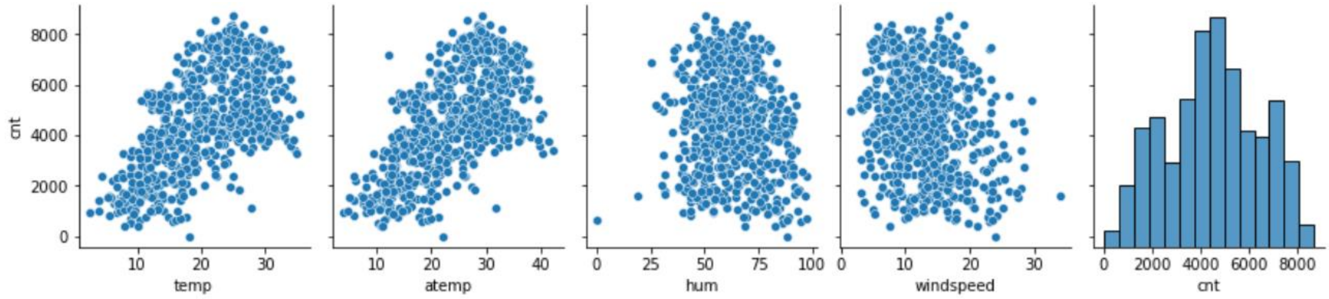
2. Why is it important to use drop first=True during dummy variable creation?

Answer:

- Dummy variables are created using one-hot encoding on non-binary categorical variables (which have more than 2 categories)
- Each value in dummy variables is either 0 or 1, here 1 signifies the presence whereas 0 signifies the absence of a particular category.
- We assume that the initial category has a value of 0 value in the first row, for example a categorical variable of 3 categories has values such as – 000, 001, 010 ,011, 100, 101, 110 and 111
- Using **drop first=True** we drop the initial category through we reduce the changes of multicollinearity while building the model.
- Thus, a categorical variable having “n” category has “n-1” dummy variables.

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

Answer: Below is the pair plot relation between numeric variables and target variable (cnt).

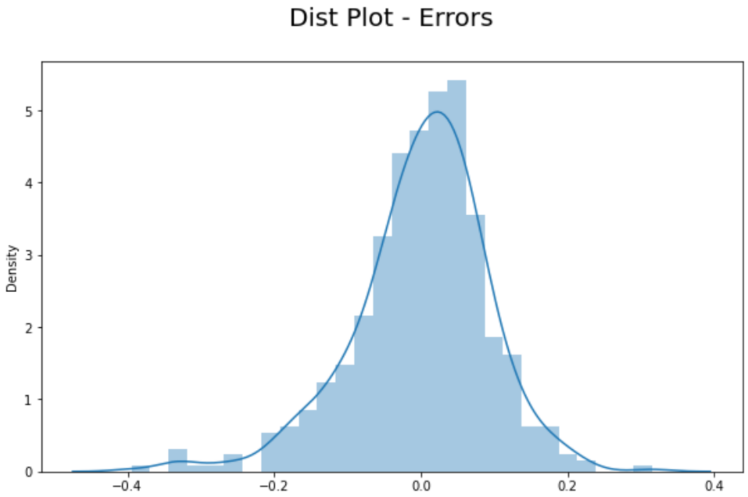


By observing the plot, we can say that numerical variables **temp** and **atemp** has a highest correlation with the target variable **cnt.**

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

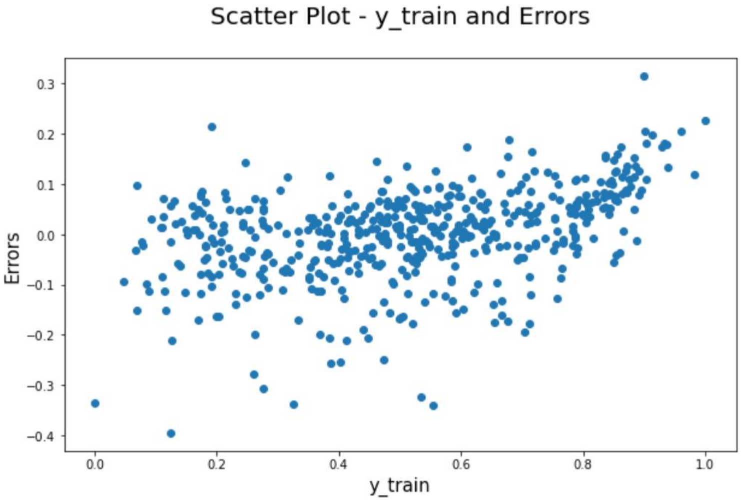
Answer: Below are the assumptions we make to validate the model on the training set:

A. The errors terms should be Normally Distributed with the mean at 0.



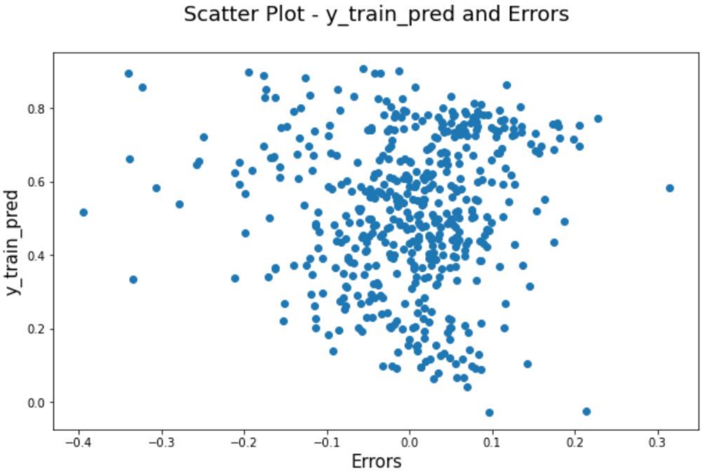
When we plot a distribution graph on error terms, we can see a **Normal Distribution** is formed which is a bell like curve with the mean at 0.

B. The errors terms should be Independent of each other.



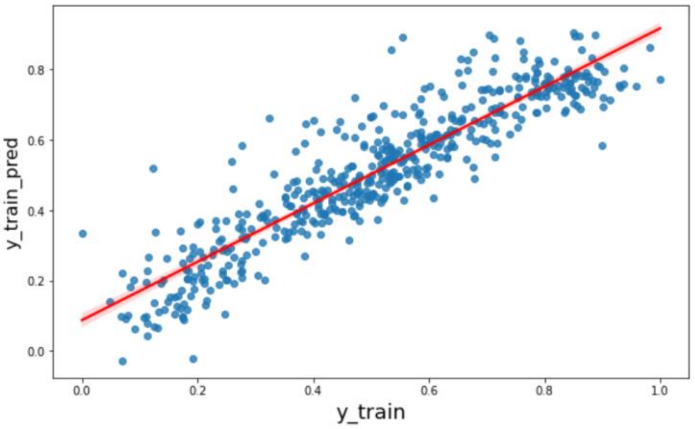
When we plot a scatter plot between the error terms and the y_train data having target variable, we can observe that there is no pattern formed between y_train and errors and he points are scattered around y = 0.

C. The error terms must have constant variance (Homoscedasticity)



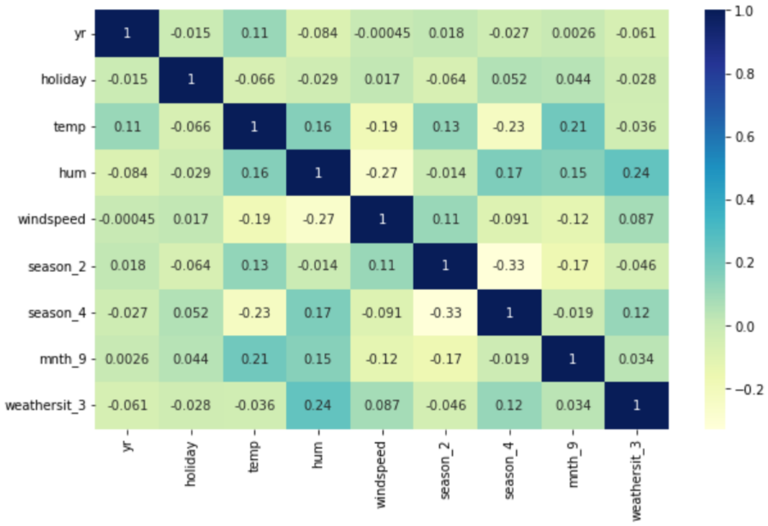
- The error terms do not form any pattern.
- The error terms have a constant variance.
- Thus, we can say the errors are **homoscedastic**

D. There must be a strong linear relationship between independent (X_train) and target variable (y_train)



When we plot a scatter plot between the actual target variable and predicted target variable, we can observe that there is a strong linear relationship between them which forms a straight line.

E. No multicollinearity between independent variables



- From the above graph we can clearly see that all the input variables are independent of each other which shows that there is No multicollinearity between independent 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 top 3 features contributing significantly towards explaining the demand of the shared bikes are :

i. **temp** - A unit increase in temperature variable **increases** the bike bookings by 0.60 units.

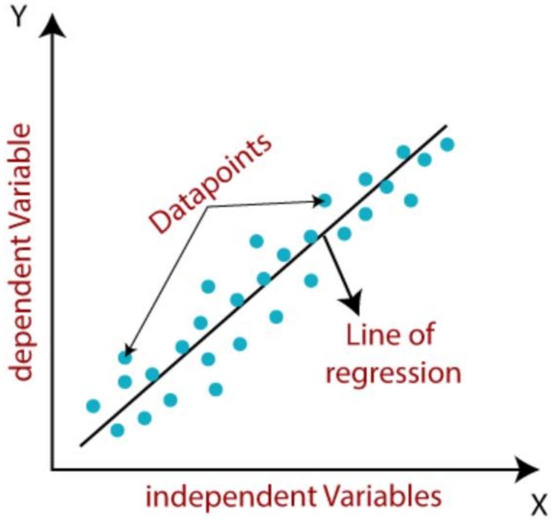
ii. **hum** - A unit increase in humidity variable **decreases** the bike bookings by 0.28 units.
iii. **yr** - A unit increase in year variable **increases** the bike bookings by 0.22 units.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

- Regression is a supervised learning technique that supports finding the correlation among variables.
- Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of our data based on some variables.
- It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or 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 (x) variables, hence called as linear regression.



- The Mathematical equation of Simple linear regression is $\boxed{y = \beta_0 + \beta_1 X + \varepsilon}$

where:

β_0 is the constant (value of Y when X = 0)

β_1 is the regression coefficient

X is the value of the independent variable

Y is the value of the independent variable and

ε is the error

- Linear regression can be further divided into two types of the algorithm:

1. **Simple Linear Regression**: When a single independent variable is used
2. **Multiple Linear regression**: When more than one independent variable is used

$$\boxed{y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \epsilon}$$

- When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized.

- **Cost function** optimizes the regression coefficients or weights and measures how a linear regression model is performing. The cost function is used to find the accuracy of the mapping function that maps the input variable to the output variable.

- While finding the best fit line we encounter that there are errors while mapping the actual values to the line. These errors are nothing but the residuals. To minimize the error squares OLS (Ordinary least square) is used.

$e_i = y_i - y_{pred}$ is provides the error for each of the data point.

• Ordinary Lease Squares method is used to minimize Residual Sum of Squares and estimate beta coefficients.

2. Explain the Anscombe's quartet in detail.

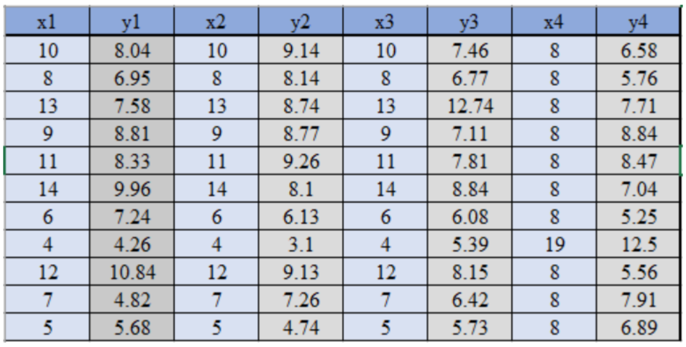
Answer:

- Anscombe's Quartet is the modal example to demonstrate the **importance of data visualization** which was developed by the statistician Francis Anscombe in 1973 to signify both the importance of plotting data before analysing it with statistical properties.

- It comprises of four data-set and each data-set consists of 11 (x,y) points.
- 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 characteristics in the dataset that fools the regression model if built.

- The basic thing to analyse 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 behaviours irrespective of statistical analysis.



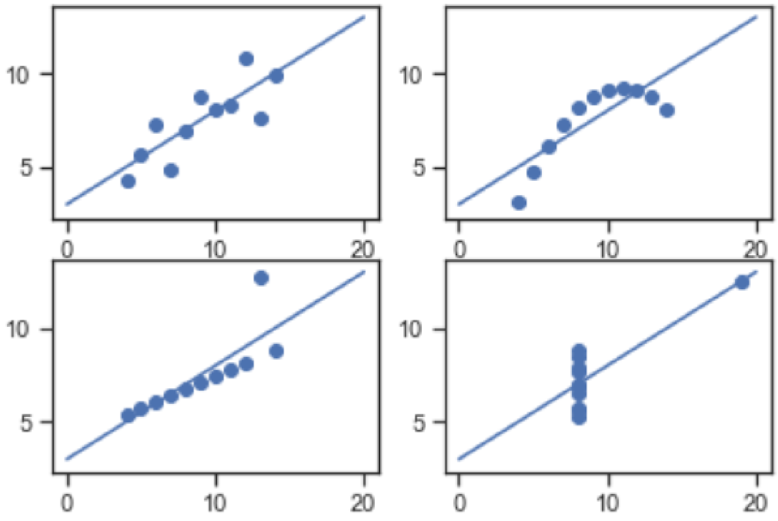
x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	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 behaviour.



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.
- Anscombe's Quartet signifies that Data-sets which are identical over a number of statistical properties, yet produce dissimilar graphs.
- This shows the importance of data visualization while exploring data.

3. What is Pearson’s R?

Answer:

- Pearson’s r is also known as **Pearson correlation coefficient** or **Pearson’s correlation coefficient.**
- It is defined in statistics as the measurement of the strength of the relationship between two variables and their association with each other.
- In simple words, Pearson’s correlation coefficient calculates the effect of change in one variable when the other variable changes.
- For example: Up till a certain age, (in most cases) a child’s height will keep increasing as his/her age increases. Of course, his/her growth depends upon various factors like genes, location, diet, lifestyle, etc.
- The correlation coefficient formula finds out the relation between the variables. It returns the values between -1 and 1.

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

Where:

  - N = the number of pairs of scores
  - Σxy = the sum of the products of paired scores
  - Σx = the sum of x scores
  - Σy = the sum of y scores
  - Σx2 = the sum of squared x scores
  - Σy2 = the sum of squared y scores

- The interpretation of the coefficients are:
    - -1 coefficient indicates strong inversely proportional relationship.
    - 0 coefficient indicates no relationship.
    - 1 coefficient indicates strong proportional relationship.

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

Answer:

What is scaling:

- Scaling is a defined as a technique to standardize the independent features present in the data set in a fixed range.
- It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

Why is scaling performed?

- When building a multiple regression model, there will be always free potential predictor variables with different range of values and selecting the just right one becomes an important exercise.
- If scaling is not done, then a machine learning algorithm tends to weigh greater values as higher and consider smaller values as the lower values, regardless of the unit of the values.
- **Example:** If an algorithm is not using the scaling method, then it can consider the value 3000 meters to be greater than 5 km but that's actually not true and, in this case, the algorithm will give wrong predictions. So, we use Scaling to bring all values to the same magnitudes and thus, tackle this issue.
- Also, if we don't do scaling some of the coefficients obtained b y fitting the regression model might be very large or small as compared to the other coefficients.

What is the difference between normalized scaling and standardized scaling?

- Scaling are of 2 types:

1. **Normalisation or Min-max Scaling -** This technique re-scales a feature or observation value with distribution value **between 0 and 1.**

$$\boxed{\frac{X-Xmin}{Xmax-Xmin}}$$

- Since values are between 0 and 1, this technique **takes cares of the outliers present in the variable**. Thus, it is also the most commonly used technique

2. **Standardization:** This technique re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$\boxed{\frac{X_i - X_{mean}}{STD}}$$

- The benefit of using standardization is that it optimises the model in background and make it much faster. When we are training a LR model, basically a Gradient Descent algorithm tries to minimise the cost function, so this minimisation becomes much faster.

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

Answer: The formula of VIF (Variance Inflation factor) is

$$\boxed{VIF = \frac{1}{1 - R^2}}$$

- Where R^2 tells how good the model is fitted and the value of R^2 lies between 0 to 1 where 1 signifies a perfectly fill model but this could be due to multicollinearity between the independent variables.
- When R square is 1 then the VIF will become infinity as 1/0 is infinity.
- Thus, the scenario of infinite VIF happens when then multicollinearity between the independent variables, that is there a very high correlation between 2 independent variables.

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

Answer:

- The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential.

- For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption.

- It’s just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.

- The Q-Q plots are useful in the linear regression to identify the train data set and test data set are from the populations with same distributions.

- The Quantile-Quantile plot is used for the following purpose:
  - Determine whether two samples are from the same population.
  - Whether two samples have the same tail
  - Whether two samples have the same distribution shape.
  - Whether two samples have common location behaviour.

- Advantages of Q-Q plot
  - Since Q-Q plot is like probability plot. So, while comparing two datasets the sample size need not to be equal.
  - Since we need to normalize the dataset, so we don’t need to care about the dimensions of values.