

## 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)

- season: Almost 32% of the bike booking were happening in season3 with a median of over 5000 booking (for the period of 2 years). This was followed by season2 & season4 with 27% & 25% of total booking.

- mnth: Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month.

- weathersit: Almost 67% of the bike booking were happening during 'weathersit1' with a median of close to 5000 booking (for the period of 2 years). This was followed by weathersit2 with 30% of total booking.

- holiday: Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is clearly biased.

- weekday: weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor.

- workingday: Almost 69% of the bike booking were happening on a 'workingday' with a median of close to 5000 booking (for the period of 2 years).

- Fall season seems to have attracted more bookings. And, in each season the booking count has increased drastically from 2018 to 2019.

- Most of the bookings have been done during the month of may, june, july, aug, sep and oct. Trend increased starting of the year till mid of the year

and then it started decreasing as we approached the end of year. Number of bookings for each month seems to have increased from 2018 to 2019.

- Clear weather attracted more bookings which seems obvious. And in comparison to previous year, i.e 2018, booking increased for each weather situation in 2019.

- Thu, Fri, Sat and Sun have more bookings as compared to the start of the week.

- When it's not holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.

- Booking seemed to be almost equal either on working day or non-working days. But, the count increased from 2018 to 2019.

- 2019 attracted more bookings from the previous year, which shows good progress in terms of business.

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

`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. It allows you to drop your first variable and identify it through all other columns being 0.

Suppose, you have a column for gender that contains 4 variables- "Male", "Female", "Other", "Unknown". So a person is either "Male", or "Female", or "Other". If they are not either of these 3, their gender is "Unknown".

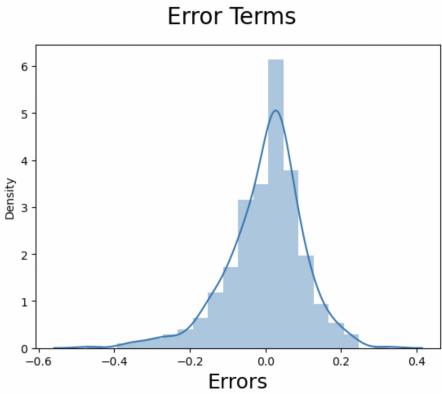
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

'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)

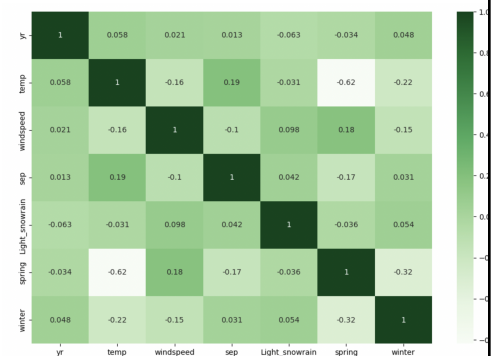
The assumptions are validated using the below:

- Check if the error terms are normally distributed
  - The error terms are normally distributed as shown in the fig below.

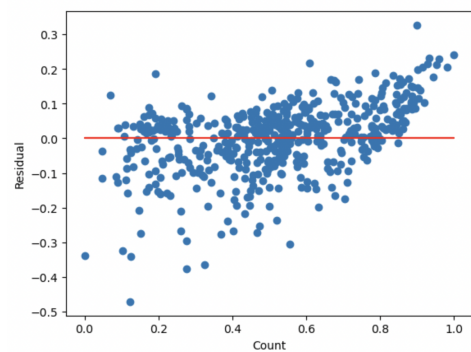
#	Assumptions	Screenshot
1.	<p>Check if error terms are normally distributed</p> <ul style="list-style-type: none"><li>- The mean lies around 0, and there is a bell shaped curve</li></ul>	

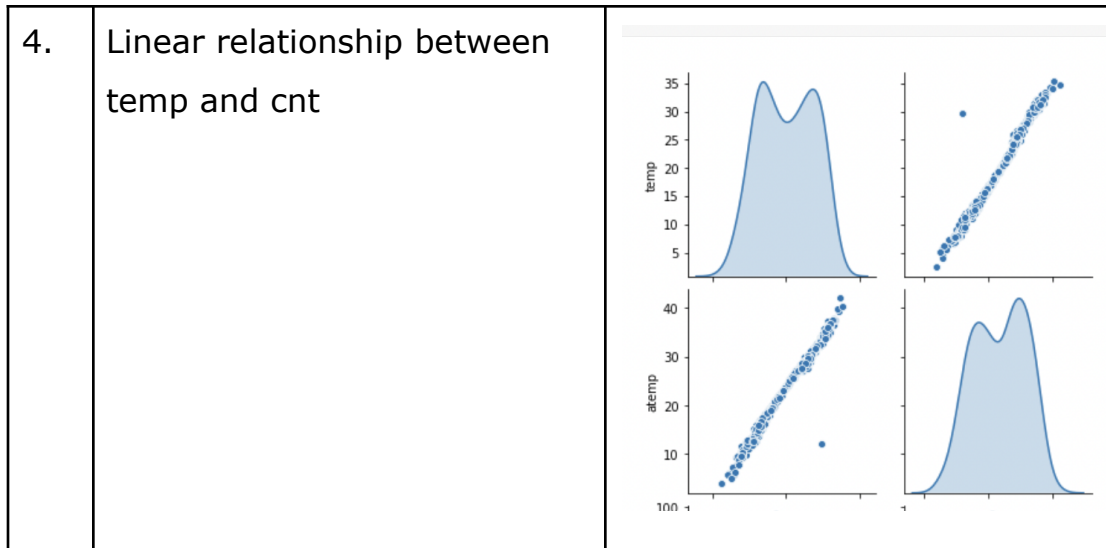
2. Check for multicollinearity
- VIF is the Variance Inflation Factor and it shows that all the values are  $< 5$
  - The heatmap doesn't show high correlation of cnt with any other variables

	Features	VIF
2	windspeed	4.68
1	temp	3.94
0	yr	2.03
5	spring	1.68
6	winter	1.32
3	sep	1.15
4	Light_snowrain	1.05



3. Check for Homoscedasticity
- There is no visible pattern that can be found. It has constant variance





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

From our equation obtained from the final model

$$\text{cnt} = 0.2276 + 0.2346 * \text{year} + 0.4374 * \text{temp} - 0.1379 * \text{windspeed} + 0.0627 * \text{sep} - 0.2812 * \text{Light\_snowrain} - 0.1126 * \text{spring} + 0.0456 * \text{winter}$$

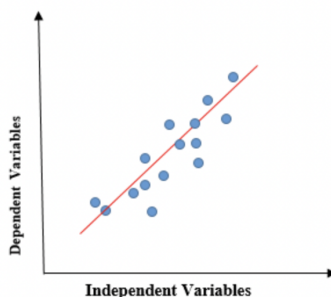
It can be said that these below variables are the top 3 features contributing significantly to the demand of shared bikes.

- Temp
- Sep
- Winter

## General Subjective Questions

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

Linear regression is a quiet and simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, *such linear regression is called multiple linear regression*. The linear regression model gives a sloped straight line describing the relationship within the variables.



The above graph presents the linear relationship between the dependent variable and independent variables. When the value of x (independent variable) increases, the value of y (dependent variable) is likewise increasing. The red line is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best.

*To calculate best-fit line linear regression uses a traditional slope-intercept form.*

$$Y = mx + c$$

Where y is the dependant variable

M is the slope

X is the independent variable

C is the intercept

A regression line can be a Positive Linear Relationship or a Negative Linear Relationship.

### **Cost function**

The cost function helps to figure out the best possible values for  $a_0$  and  $a_1$ , which provides the best fit line for the data points.

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. This mapping function is also known as the Hypothesis function.

In Linear Regression, Mean Squared Error (MSE) cost function is used, which is the average of squared error that occurred between the predicted values and actual values.

By simple linear equation  $y=mx+b$  we can calculate MSE as:

Let's  $y$  = actual values,  $y_i$  = predicted values

$$MSE = \frac{1}{N} \sum_{i=1}^n (y_i - (mx_i + b))^2$$

Using the MSE function, we will change the values of  $a_0$  and  $a_1$  such that the MSE value settles at the minima. Model parameters  $x_i$ ,  $b$  ( $a_0, a_1$ ) can be manipulated to minimize the cost function. These parameters can be determined using the gradient descent method so that the cost function value is minimum.

## Gradient descent

Gradient descent is a method of updating  $a_0$  and  $a_1$  to minimize the cost function (MSE). A regression model uses gradient descent to update the coefficients of the line ( $a_0, a_1 \Rightarrow x_i, b$ ) by reducing the cost function by a random selection of coefficient values and then iteratively update the values to reach the minimum cost function.

## Assumptions of simple linear regression

- There is a *linear relationship* between X and Y
- Error terms are *normally distributed* with mean zero(not X, Y)
- Error terms are *independent* of each other
- Error terms have *constant variance* (homoscedasticity)

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

**Anscombe's quartet** is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another.

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data.

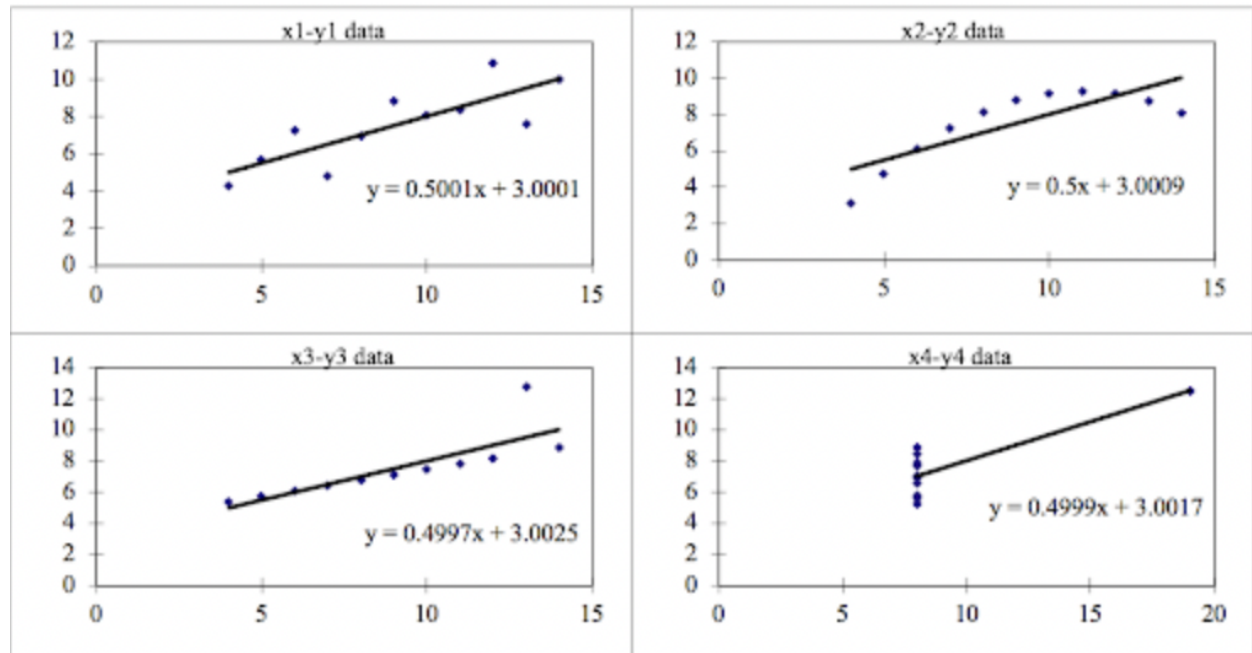


Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		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	

From the data above you can see that:

- Mean is the same for all 4 sets of data for both y and x which is 7.5 and 9 respectively
- SD is the same for all 4 sets. Which is 3.16.
- When calculated the slope, intercept, variance its all the same.

But when you scatter plot them you can see that they are very different .



Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone

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

The Pearson correlation coefficient ( $r$ ) is the most common way of measuring a linear correlation. It is a number between  $-1$  and  $1$  that measures the strength and direction of the relationship between two variables.

Pearson correlation coefficient ( $r$ )	Correlation type	Interpretation	Example
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the same direction.	<p>Baby length &amp; weight:</p> <p>The longer the baby, the heavier their weight.</p>
0	No correlation	There is no relationship between the variables.	<p>Car price &amp; width of windshield wipers:</p> <p>The price of a car is not related to the width of its windshield wipers.</p>

Between  0 and -1	Negative correlation	When one variable changes, the other variable changes in the opposite direction.	Elevation & air pressure:  The higher the elevation, the lower the air pressure.
-------------------------	-------------------------	---	--

The Pearson correlation coefficient ( $r$ ) is the most widely used correlation coefficient and is known by many names:

- Pearson's  $r$
- 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.

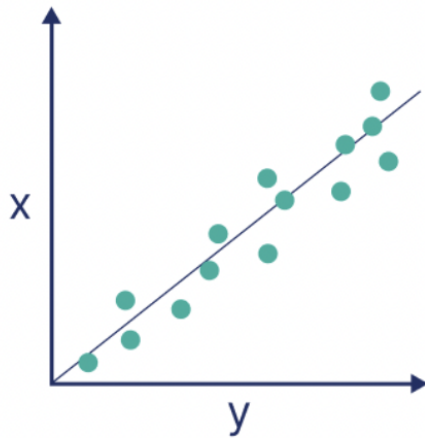
<b>Pearson correlation coefficient (<math>r</math>) value</b>	<b>Strength</b>	<b>Direction</b>
Greater than .5	Strong	Positive

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

The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.

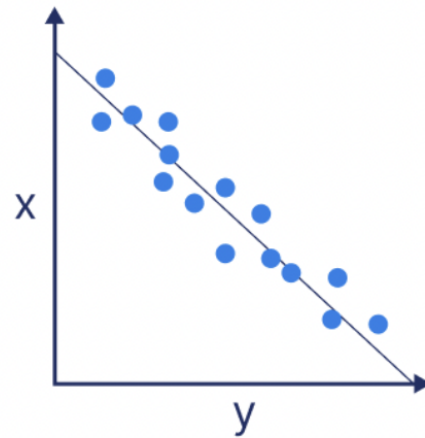
### Strong positive correlation

$$r > .5$$



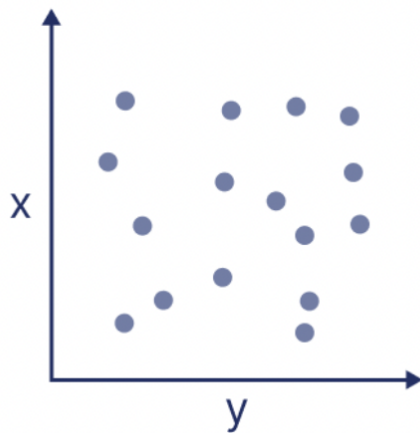
### Strong negative correlation

$$r < -.5$$



### No correlation

$$r = 0$$



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

### **What?**

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.

### **Why?**

Most of the time, the collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then the algorithm only takes magnitude in account and not units hence incorrect modeling. 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:

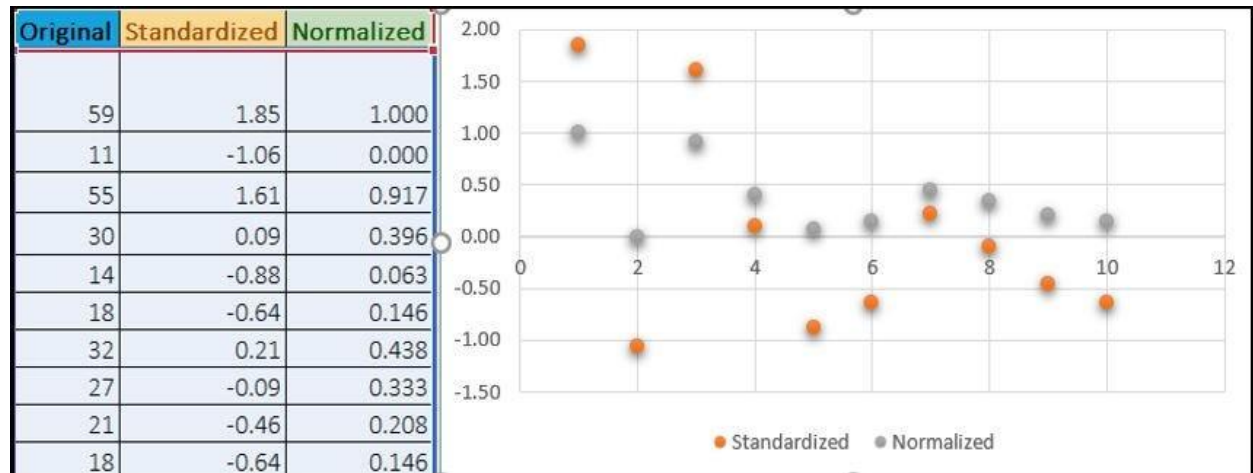
- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean ( $\mu$ ) zero and standard deviation one ( $\sigma$ ).

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

- `sklearn.preprocessing.scale` helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

**Example:** Below shows examples of Standardized and Normalized scaling on original values.





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  $VIF = \text{infinity}$ . 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. The value of VIF is calculated by the below formula:

$$VIF_i = \frac{1}{1 - R_i^2}$$

Where, 'i' refers to the ith variable.

If the R-squared value is equal to 1 then the denominator of the above formula becomes 0 and the overall value becomes infinite. It denotes perfect correlation in variables.

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

Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against. For example, if you are testing if the distribution of age of employees in your team is normally distributed, you are comparing the quantiles of your team members' age vs quantile from a normally distributed curve. If two quantiles are sampled from the same distribution, they should roughly fall in a straight line.

Since this is a visual tool for comparison, results can also be quite subjective nonetheless useful in the understanding underlying distribution of a variable(s)

Q-Q plot can also be used to test distribution amongst 2 different datasets. For example, if dataset 1, the age variable has 200 records and dataset 2, the age variable has 20 records, it is possible to compare the distributions of these datasets to see if they are indeed the same. This can be particularly helpful in machine learning, where we split data into train-validation-test to see if the distribution is indeed the same. It is also used in the post-deployment scenarios to identify covariate shift/dataset shift/concept shift visually.