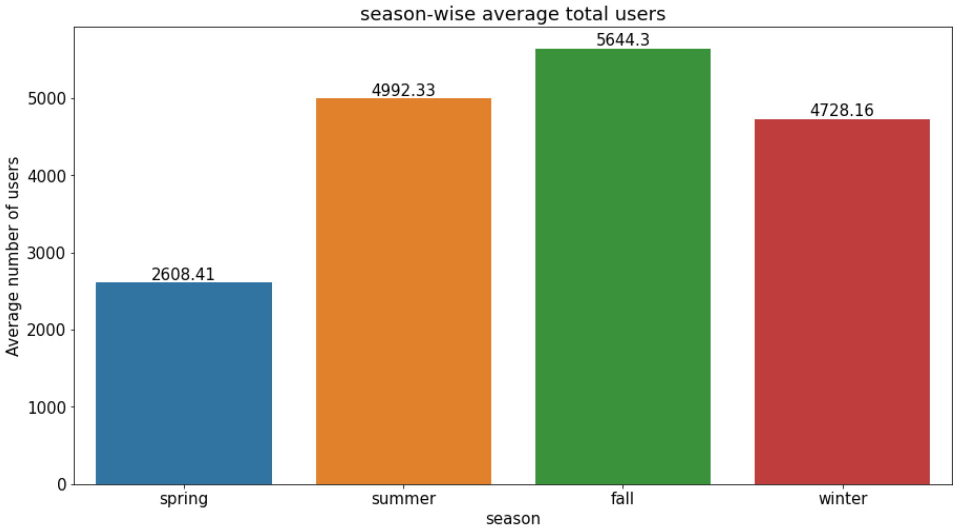
Assignment-based Subjective Questions

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

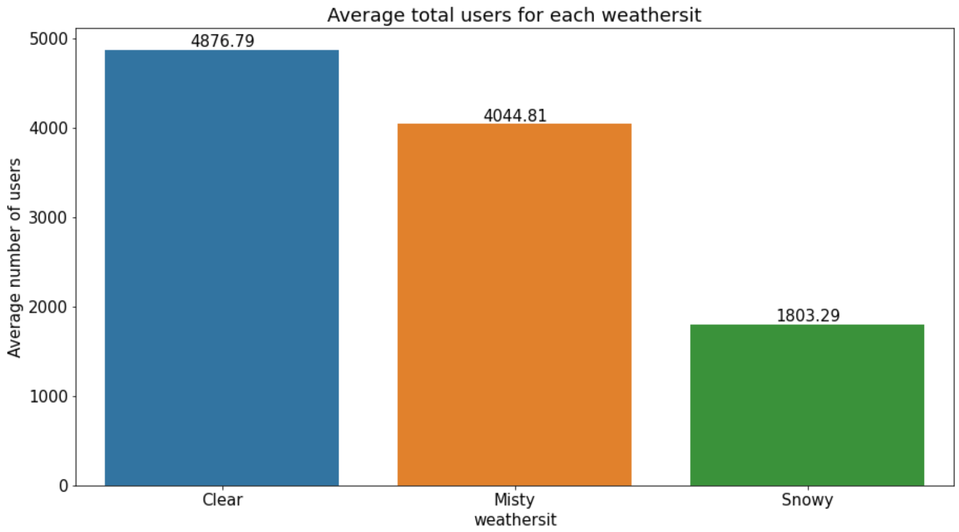
Some of the categorical variables showed noticeable differences in the pattern of demand for bikes. These variables are discussed below.

“Season” Variable:



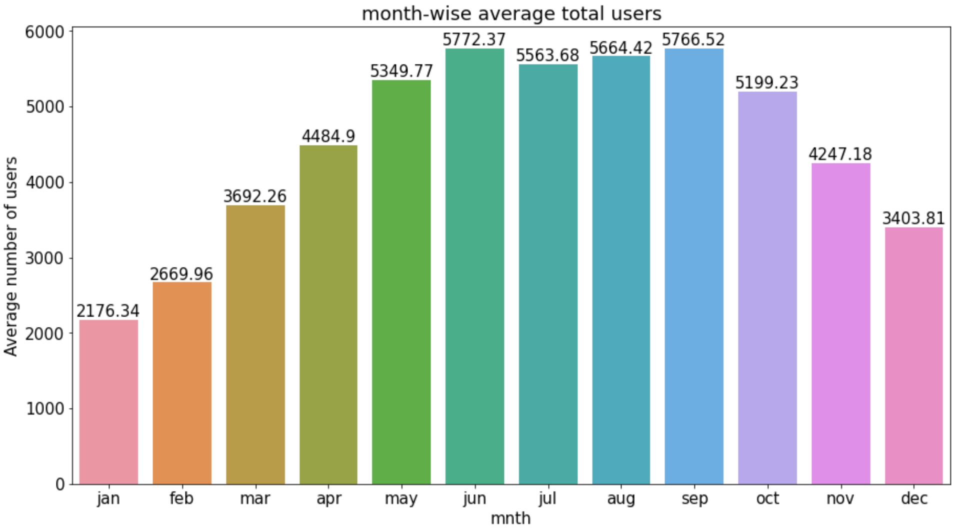
It can be observed that people tend to rent bikes during the fall season more. Also, the demand is surprisingly low during the spring season.

“Weathersit” Variable:



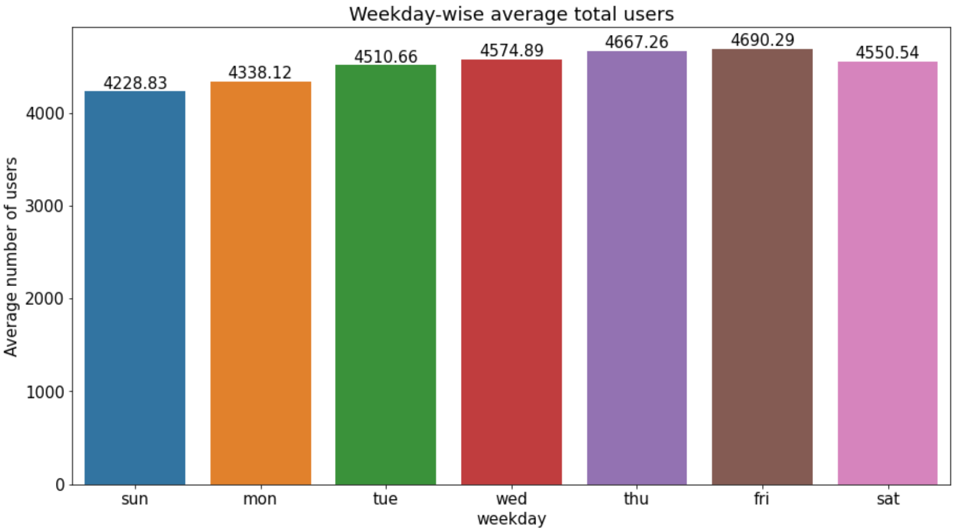
It could be observed that the number of users is high during clear weather and the demand is low when the weather is snowy.

“Month” Variable:



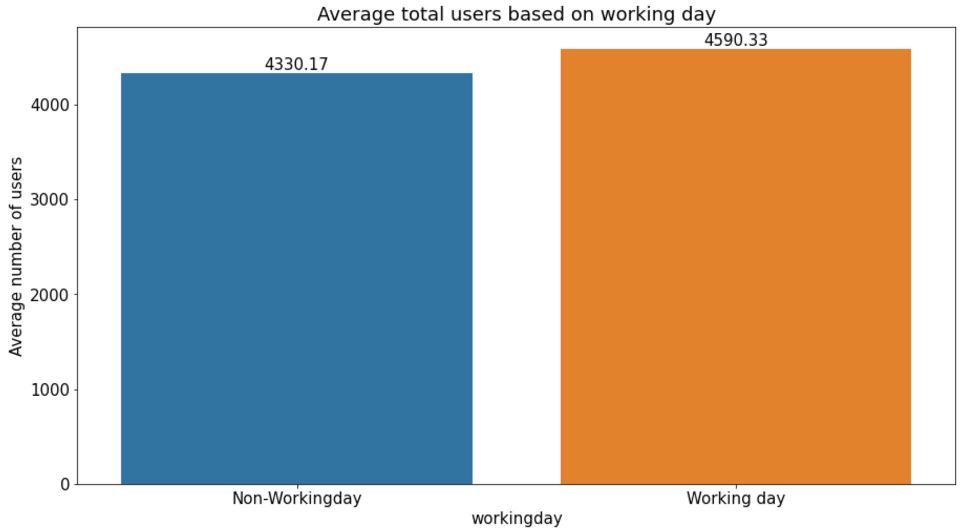
From the above graph, it could be noticed that the number of users is high during the months of June and September. Also, the number of users decreased during the December, January and February months.

“Weekday” Variable:



The number of users is high on Fridays and is the lowest on Sundays.

“Workingday” Variable:



The number of users is high on working days when compared to non-working days.

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

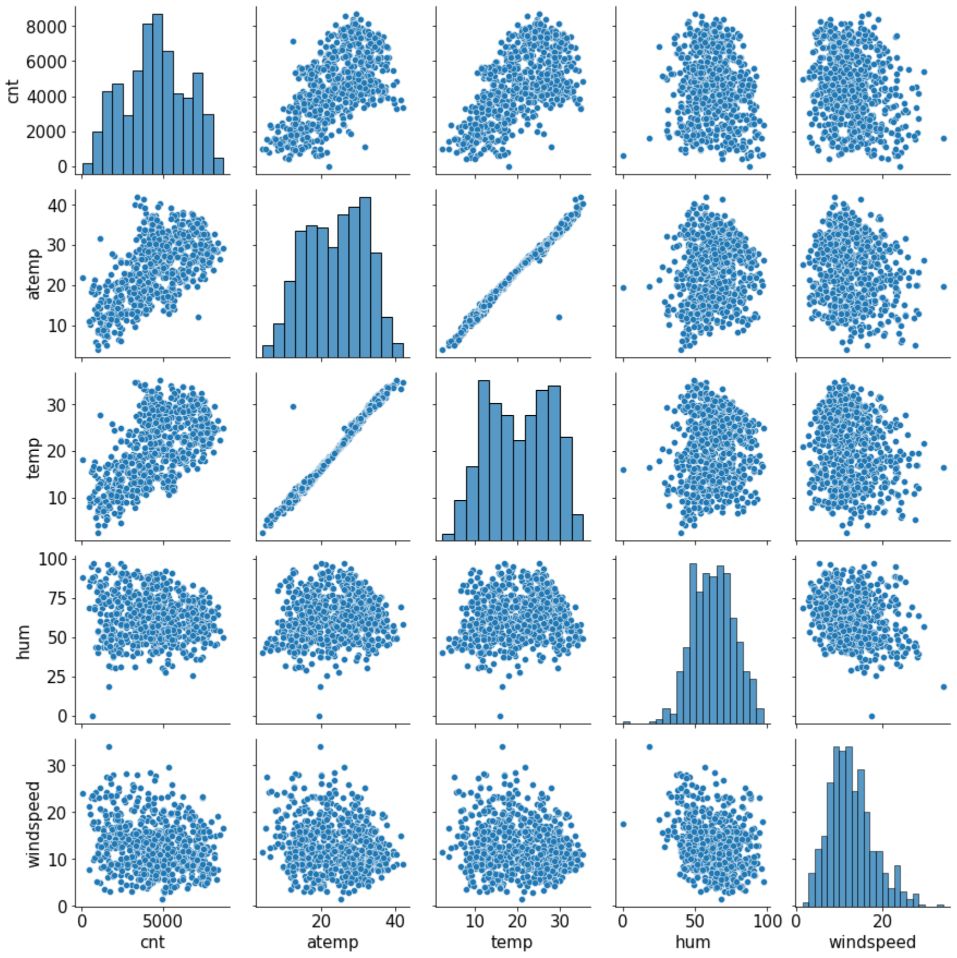
For a categorical variable with n variables, the number of dummy variables required to represent all the categories would be n-1. During dummy variable creation in pandas, we use drop_first=True to eliminate the first column of one-hot encoding matrix.

For example, if there are 4 different categories in a categorical variable, by creating dummy variables for each of the categories, we would end up with 4 columns with each column having 0 and 1 as possible values. However, we would not need all 4 columns to represent the different categories as the categories can be encoded with just three variables. This can be seen in the following table.

Without drop_first=True	***With drop_first=True***
Category 1 – 1000	*Category 1 – 000*
Category 2 - 0100	*Category 2 - 100*
Category 3 - 0010	*Category 3 - 010*
Category 4 - 0001	*Category 4 - 001*

Also, by setting the drop_first argument as True, the correlation among the dummy variables will be reduced.

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



From the pair-plot of numerical variables shown above, we could see that atemp and temp have more correlation with the target variable “cnt”.

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

After building the model. I validated the assumptions of linear regression. The different assumptions and the way they are validated are as follows.

Assumption 1 – Linear relationship between X and Y:

I have assessed the linear relationship between the independent and dependent variables using the statistical parameters provided by statsmodels library. These parameters include P-value, F-statistic and R-squared values. These parameters are checked to see if they are within the acceptable range.

Assumption 2 – Normality of error terms:

The error terms are derived by getting the difference between the predicted values and the ground truth target values. The error terms are then plotted on a distribution plot. The resulting graph showed a normal distribution.

Assumption 3 – Error terms are independent of each other:

The error terms vs predicted values are plotted using regplot and the distribution of values in the plot is checked. The plot is checked for any patterns. When there is no pattern seen in the plot, it is determined that the error terms are independent of each other.

Assumption 4 – Error terms have constant variance (homoscedasticity)

When the ground truth values are plotted against predicted values from the model using a regplot, we should be seeing evenly spread-out data points in the plot along the regression line. This confirms the homoscedasticity of the model.

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

The top three features contributing significantly towards explaining the demand for shared bikes are:

- Atemp – the feeling temperature.
- Year
- The season “winter”

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Linear regression is a predictive modelling technique that builds a mathematical linear relationship between a dependent variable (target value) with one or more independent variables. In other words, a linear regression model helps predict a target value when provided with a set of independent variables which have influence on the target value.

When there is only one independent variable used for the model, it is called a simple linear regression. When there is more than one independent variable used, it is called multiple linear regression.

The mathematical representation of linear regression is as follows:

$$y = \beta_0 x_0 + \beta_1 x_1 + \ \dots \ + \beta_n x_n + c$$

Where,

y , is the target variable

$x_{0 \dots n}$, are the independent variables

$\beta_{0 \dots n}$, are the coefficients (slope) for each of the independent variables

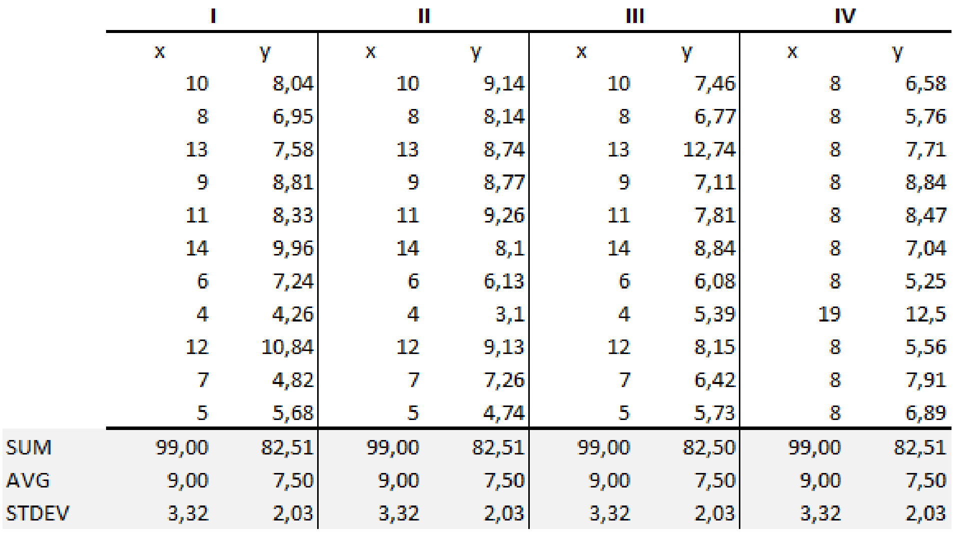
c , is the intercept pointing where the linear regression line meets the y-axis

Linear regression can be used in predictive analysis of a continuous variable target value. Examples include predicting annual sales of a company, prediction of stock prices, prediction of house price, etc., While modelling a linear regression model, the objective is to find a linear equation that can best determine the value of target variable when provided with one or more independent variables.

2. **Explain the Anscombe's quartet in detail.**

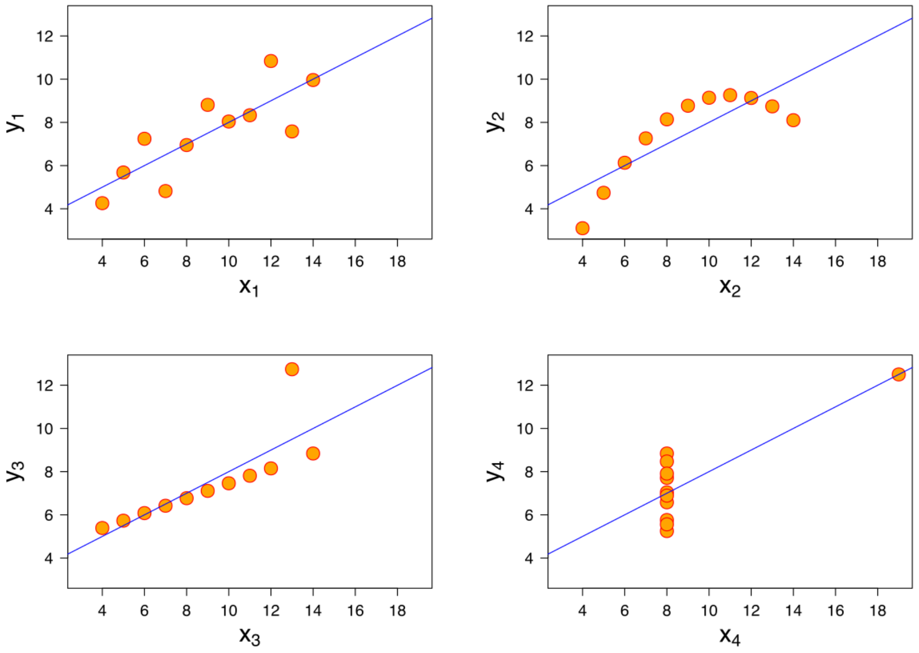
Anscombe's quartet comprises of 4 different datasets which have identical descriptive statistics namely mean, sum, standard deviation. Though, these 4 different datasets have similar statistics mathematically, things change when the values in the dataset are plotted on a graph using scatter plot.

The four datasets are represented in the below table



	I		II		III		IV	
	x	y	x	y	x	y	x	y
	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
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

It could be seen that the summary statistics look identical. When plotting these 4 datasets on a graph, we get the below distributions.



The idea of these 4 datasets is to illustrate the importance of plotting the graphs during data analysis.

In the four plots shown above, only the first dataset with x1 and y1 data can be fit using a linear regression. The other three datasets cannot handle linear regression as they do not show linear relationship, or they have high correlation coefficient, or they contain outliers.

3. What is Pearson's R

Pearson's R is a statistical measurement that shows the strength of linear association between two continuous variables. Pearson's R can take a value between –1 and 1.

- When the Pearson's R is positive, it shows a positive correlation between the two variables. Meaning, when one variable goes up the other variable goes up too.
- When the Pearson's R is negative, it shows a negative correlation between the two variables. Meaning, when one variable goes up the other variable goes down.
- When the value is 0, it shows there is no relationship between the two variables.
- When the value is –1 or 1, it means the correlation between the two variables is perfectly equal.

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

Scaling is a process of bringing the ranges of different variables to a common range.

For example, in a house price dataset, the range of values in the "area of the house" variable could be 0 to 100 metres. But the range of values in the "price of the house" variable could be 0 to ten million. Handling two different variables with completely different ranges would increase the time taken for the model to converge. Also, not scaling the distributions would result in large coefficients which could undermine some independent variables thereby resulting in an underfit model.

Normalized scaling and standardized scaling are two types of scaling that are widely used by data scientists. The table below explains the differences between the two.

Normalized Scaling	**Standardized Scaling**
The minimum and maximum value of the variable is used to scale the data.	The mean and standard deviation is used for scaling the data.
This is used when the features have different ranges.	This is used to ensure zero mean and 1 as standard deviation for a given data.
The scaling is affected by outliers as the scaler might pick these outliers as minimum or maximum values.	Outliers do not have an effect.
This can be applied for any distribution	This can be applied when the distribution is normal.
The scaled values can be in the range of (0, 1) or (-1, 1)	No limits on the scaled values

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

Variance Inflation Factor or VIF is the measure of correlation between two independent variables in a model. This can be used as a measure of multicollinearity when building the model.

The formula for VIF is $VIF = \frac{1}{1 - R^2}$. When the R-squared value is 1, the value of VIF goes to infinity. R-squared value of 1 means the variables are correlated perfectly. This means one or more of the variables used can be unnecessary for a model's performance. Variables with VIF value of infinity can be removed from the model without impacting the performance of the model.

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

Q-Q plot or Quantile-Quantile plot is a probability plot which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. This helps in determining if two datasets come from population with a common distribution.

Uses of Q-Q plot:

A Q-Q plot is used to compare the shapes of distribution with which you could infer location, scale and skewness of two different distributions. This can help in detecting outliers, change in scale, symmetry, etc.,

Importance of Q-Q plot:

During model building with two data samples, it is desirable to know if an assumption of common distribution is justified. If the two samples have a common distribution, the location and scale estimators can pool both data sets to obtain common location and scale. In another case, if the distribution of two samples differ, Q-Q plot would help understand the differences. This can also provide more insights into the characteristics of the difference compared to other common analytical methods.