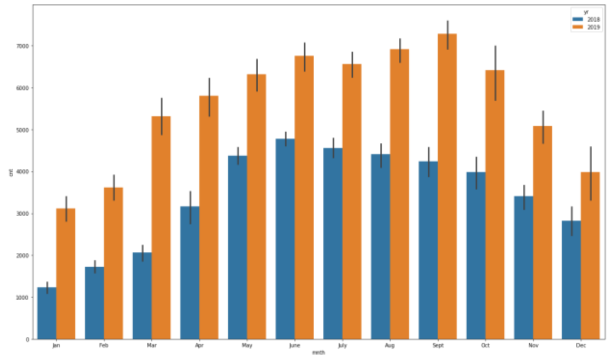
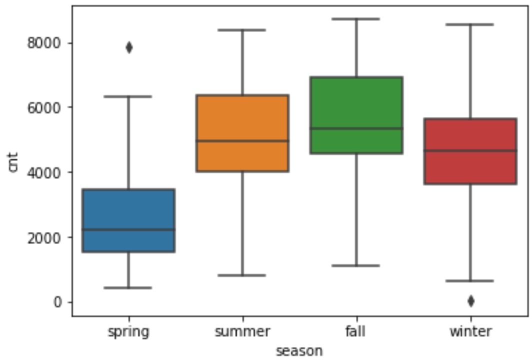
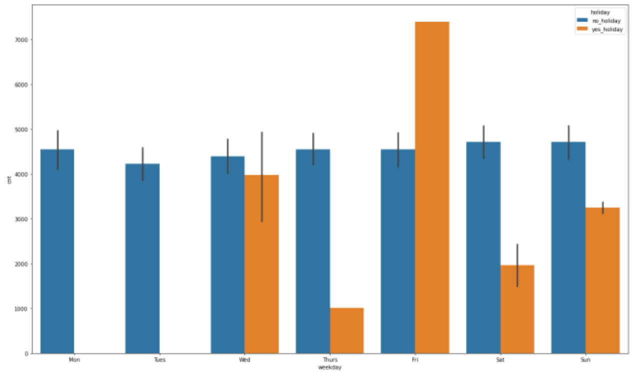
Assignment-based Subjective Questions

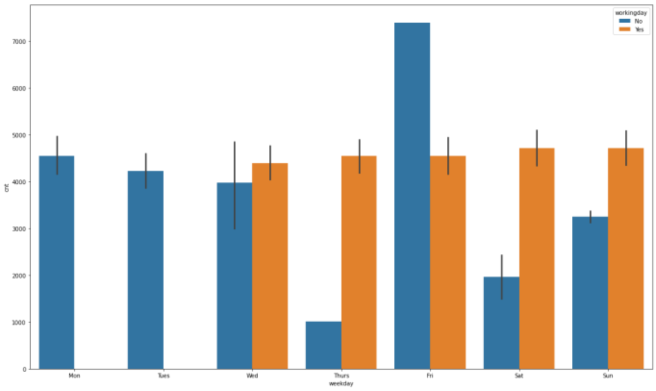
Q 1) From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

A 1) By the following plots I analyse that,









The best season where people preferred to try/use the service more is **fall Season** as the season which comes in **September month,** at that time the **weather is almost clear** or with **few clouds** in the sky & as **Friday is the day-off or holiday** that's why people came out more and as at that time is more pleasant and more favourable/comfortable for the people to try service. Also people are engaging more every preceding year. Other than that we can see **similar hikes on everyday** as that comes **from those who use the service on daily basic** for work. As weather is not something we can control so, the users spread splits as according to the current weather, if the weather is clear, People will use the service more.

Q 2) Why is it important to use **drop_first=True** during dummy variable creation?

A 2) It important to use **drop_first=True** during dummy variable creation because it makes a similar but efficient interpretation about the assignment of the variables values as binary encoded entry to the system. Which makes processing faster and we don't lose any data from our dataset as any variable with like, (0,0) have its own presence in the data somehow. Which we try to create by using dummy variables. That's why we use **drop_first= True** . So, that whatever the variable is assigned on the first place while creating dummy will be removed.

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

A 3) Temperature had the highest correlation with the target variable. As we can see or confirm it with heatmap as well.

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

A 4) There were a lot of variables which came into picture after creating dummies. But the problem was that, the independent variables were also somehow correlated with each other like, Humidity, Temperature, Season, Weekdays, workday, etc. so, I used the filtration for what to choose and what not by considering P-value and VIF. Here I rejected or dropped those variables which had extremely high P-value like, October month which had very high P-value with very low VIF, so it wasn't significant to work with those kind of variables and after removing variables by P-value (which were higher than 0.005). I went with VIF, removed variables which had more than 5 as VIF value and did the same after which I came up with 13 independent variables which were in the safe zone as we wanted. So, that's how I validate the assumptions of Linear Regression after building the model on the training set.

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

A 5) Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes are :-

1) Year, as year preceding the number of people will increases as well to use the service.
2) Wind speed, as wind speed increases the number of user decreases.
3) Working day, more the people work or we approach close to those people where people find easy to get in touch with our service and their work, more we will get the users for our service.

General Subjective Questions

Q 1) Explain the linear regression algorithm in detail.

A 1) The Basic Equation of linear regression,

$$y= c+m_1x_1+m_2x_2+...+m_nx_n\mathrm{y}$$

Where,

- y is the response
- c is the intercept
- m_1 is the coefficient for the first feature
- m_n is the coefficient for the nth feature

and to learn a linear regression model basically means that we born the algorithm to learn all of these coefficient including the 'c' constant or the intercept.

Steps for Linear Regression :-

- **Create X & Y** (X is /are independent variables & y is dependent variable)
  In any predictive analytics we always want to train model on the train dataset & you want to validate the model on the unseen hidden dataset k/as validation or test dataset. So, the terminology we use are train & test a dataset.

- **Create train & test sets** (typically used ratios are 70:30, 80:20)
  These ratios mean that, on 70% we'll train the model and on 30% we will test the set on which you'll evaluate the model using techniques like, (R^2) and others.

So, once we create a training set & the test set.

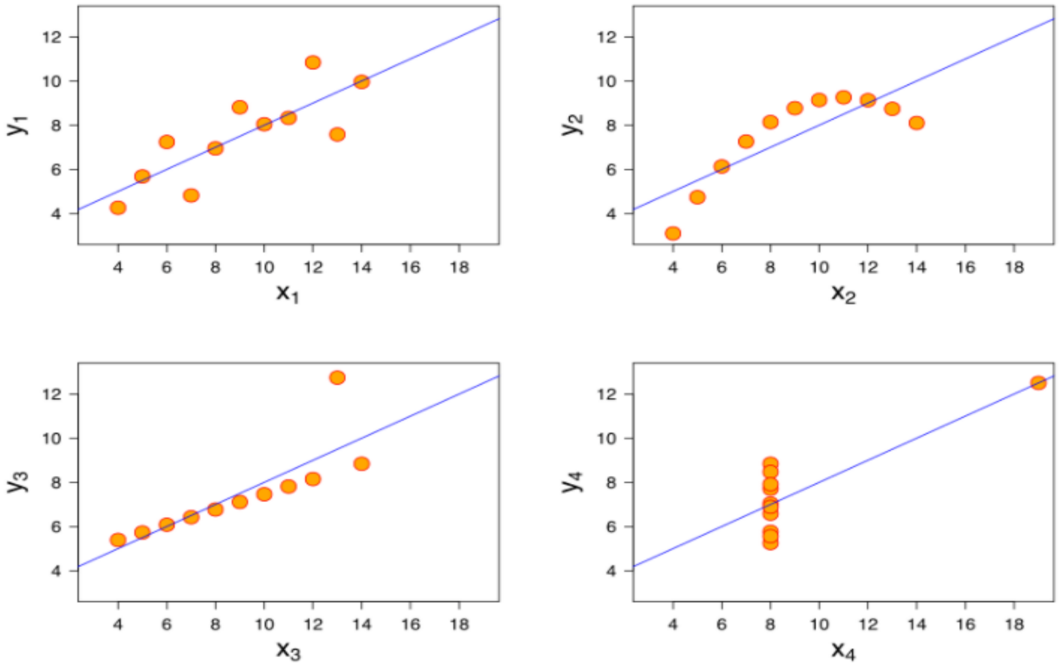
- **train a model on the training set** (i.e., basically means to learn the coefficients)
- **Evaluate the model** on test set. (also can evaluate on training set)
  It should definitely fit very well or generalise well to the test set.

- **Make Predictions** on the test set
  Basically prediction about the best fit line and the interpretation on the coefficients we got.

This is the linear regression algorithm which we can follow to build a decent model for good predictions.

Q 2) Explain the Anscombe's quartet in detail.

A 2) Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics but yet have very different distribution and appearance when graphed.



Like, in the given graph. The values which we used to plot the function gives similar results which we can analyse with the graphical representation like, by scatter plot, bar plot, etc. These kind of special cases where data interpret similar result statistically but have different structure when it comes to visualization are k/as **Anscombe's quartet** .

Q 3) What is Pearson's R?

A 3) Correlation coefficients are used to measure how strong a relationship is between two variables. There are several types of correlation coefficient, but the most popular is Pearson's. **Pearson's correlation** (also called Pearson's *R*) is a **correlation coefficient** commonly used in linear regression. Which we calculate by following formula:-

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

By this formula, we get 3 kind of interpretations which means that,

- If a correlation coefficient is positive(+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.
- If a correlation coefficient is negative (-1), means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other.
- If Zero, means that for every increase, there isn't a positive or negative increase. The two just aren't related.

The absolute value of the correlation coefficient gives us the relationship strength. The larger the number, the stronger the relationship.
For example, |-0.70| = 0.70, which has a stronger relationship than 0.60.

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

A 4) Scaling in a way by virtue of which we try to put the values in favourable range. Scaling is performed to scale the different quantities with different values like some variables have values in tens, some have in thousands. So, to see all the values in a same range so, to correlate one variable with the other variables. The simple difference between Normalized Scaling & standardized scaling is that, normalized scaling is a technique where we scale the values in between 0 & 1. On the other hand, Standardized scaling is a way where we scale the variables in a set range which can be any range like, (-1 to 1), (-3 to 3) etc. with some standard variance. Which allow us to get visuals of very large values in a short reasonable scale. People uses Normalized Scaling more because, Values are not allowed to spread out of 0 & 1 in Normalized Scaling so, it cares about outliers very well as it keep the outliers very close to 0 or 1.

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

A 5) 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, we need to drop one/some of the variables from the dataset which is causing this perfect multi-collinearity.

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

A 6) (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot. If the two data sets come from a common distribution, the points will fall on that reference line.