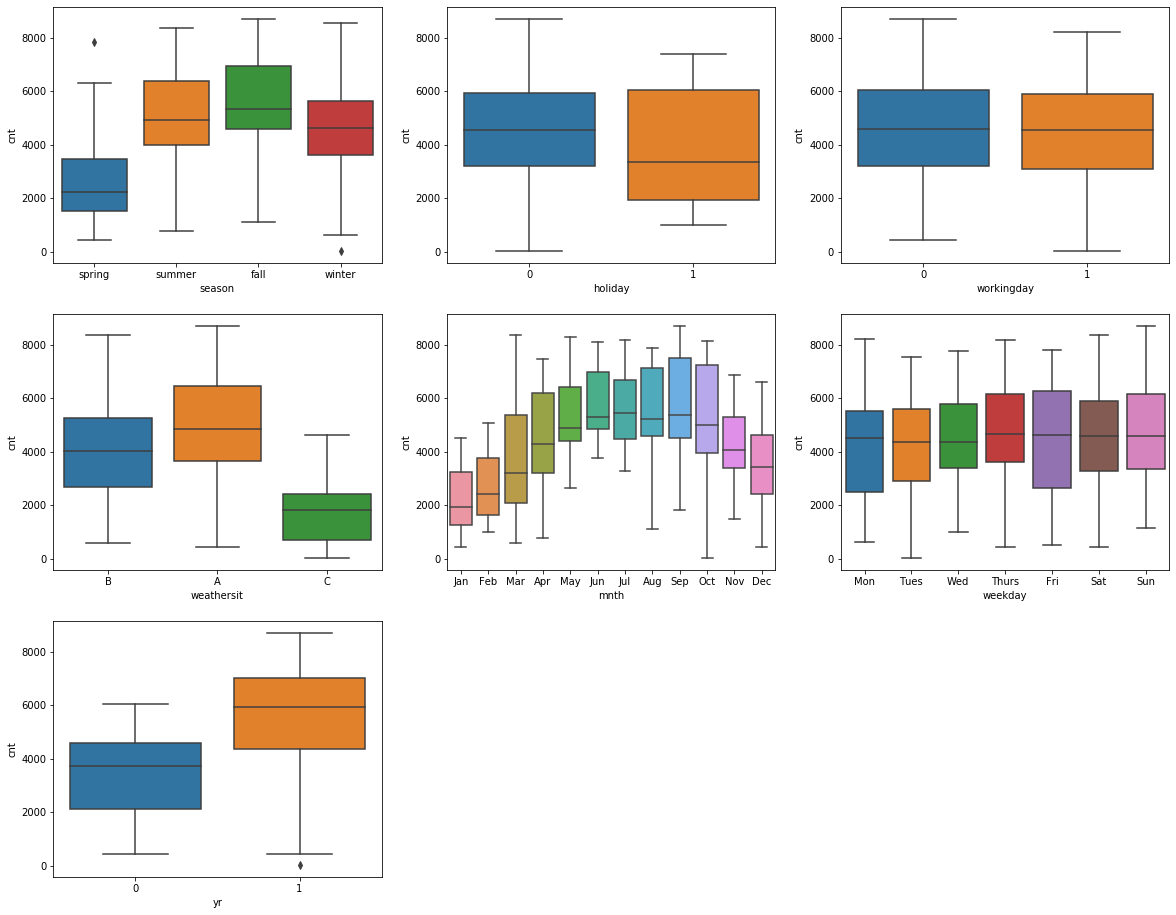
**Assignment-based Subjective Questions**

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

Ans. If we check in seasons in plot. We get the information that no of demand of shared bikes is more when the season is fall and weather is clear. If we specifically check about the months then aug, sep and oct has the best demands.

Please refer fig below:



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

Ans. The (drop\_first=True) is used to drop the first new column that we get after encoding, so we get less created variables.

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

Ans. Temp variable

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

Ans. By:

1. Linear relationship between X and Y.
2. Multicollinearity: This phenomenon exists when the independent variables are found to be moderately or highly correlated. In a model with correlated variables, it becomes a tough task to figure out the true relationship of a predictors with response variable. In other words, it becomes difficult to find out which variable is contributing to predict the response variable.
3. Normal Distribution of error terms: If the error terms are non- normally distributed, confidence intervals may become too wide or narrow. Once confidence interval becomes unstable, it leads to difficulty in estimating coefficients based on minimization of least squares. Presence of non – normal distribution suggests that there are a few unusual data points which must be studied closely to make a better model.

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

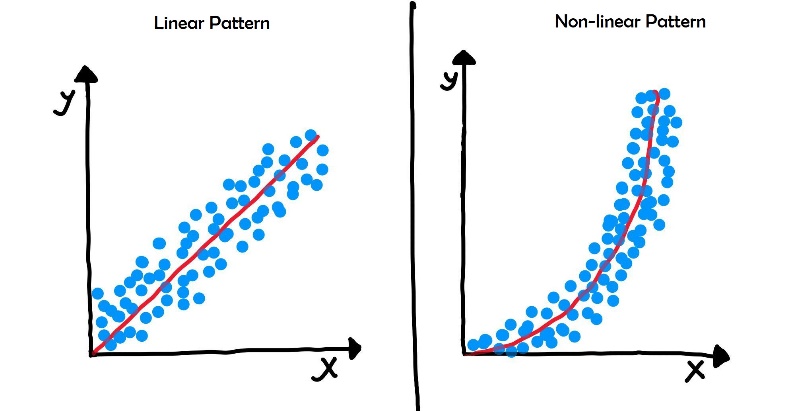
Ans. The top 3 features contributing significantly towards explaining the demand of the shared bikes are:

1. Temp is highly correlated with the cnt variable.
2. Temp, windspeed, yr, season\_spring, mnth\_july, weathersit\_C are having less VIF than 5 and even they have p > |t| is 0.

**General Subjective Questions**

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

Ans. LinearRegression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

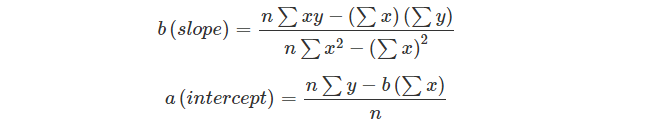


Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.  
In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

Mathematically, we can write a linear regression equation as:

Y = a + bx

Where a and b given by the formulas:



Here, x and y are two variables on the regression line.

b = Slope of the line

a = y-intercept of the line.

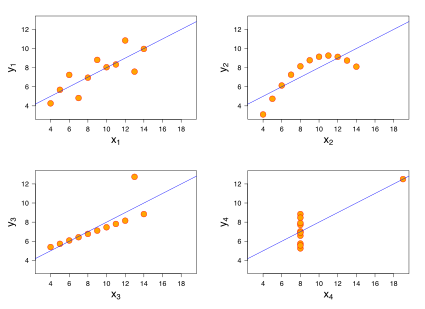
x = Independent variable from dataset

y = Dependent variable from dataset

Use Cases of Linear Regression:

1. Prediction of trends and Sales targets – To predict how industry is performing or how many sales targets industry may achieve in the future.
2. Price Prediction – Using regression to predict the change in price of stock or product.
3. Risk Management- Using regression to the analysis of Risk Management in the financial and insurance sector.
4. **Explain the Anscombe’s quartet in detail.**

Ans. [Anscombe’s quartet](http://en.wikipedia.org/wiki/Anscombe%27s_quartet) comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each [dataset](http://en.wikipedia.org/wiki/Data_set) consists of eleven (x,y) points. They were constructed in 1973 by the statistician [Francis Anscombe](http://en.wikipedia.org/wiki/Francis_Anscombe) to demonstrate both the importance of graphing data before analysing it and the effect of [outliers](http://en.wikipedia.org/wiki/Outlier) on statistical properties.



* The first [scatter plot](https://en.wikipedia.org/wiki/Scatter_plot) (top left) appears to be a simple linear relationship, corresponding to two [variables](https://en.wikipedia.org/wiki/Variable_(mathematics)) correlated where y could be modelled as [gaussian](https://en.wikipedia.org/wiki/Normal_distribution) with mean linearly dependent on x.
* The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the [Pearson correlation coefficient](https://en.wikipedia.org/wiki/Pearson_correlation_coefficient) is not relevant. A more general regression and the corresponding [coefficient of determination](https://en.wikipedia.org/wiki/Coefficient_of_determination) would be more appropriate.
* In the third graph (bottom left), the distribution is linear, but should have a different [regression line](https://en.wikipedia.org/wiki/Regression_line) (a [robust regression](https://en.wikipedia.org/wiki/Robust_regression) would have been called for). The calculated regression is offset by the one [outlier](https://en.wikipedia.org/wiki/Outlier) which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
* Finally, the fourth graph (bottom right) shows an example when one [high-leverage point](https://en.wikipedia.org/wiki/High-leverage_point) is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

1. **What is Pearson’s R?**

Ans. Correlation between sets of data is a measure of how well they are related. The most common measure of correlation in stats is the Pearson Correlation. The full name is the **Pearson Product Moment Correlation (PPMC)**. It shows the [linear relationship](https://www.statisticshowto.com/probability-and-statistics/regression-analysis/find-a-linear-regression-equation/) between two sets of data. In simple terms, it answers the question, *Can I draw a line graph to represent the data?* Two letters are used to represent the Pearson correlation: Greek letter rho (ρ) for a population and the letter “r” for a sample.

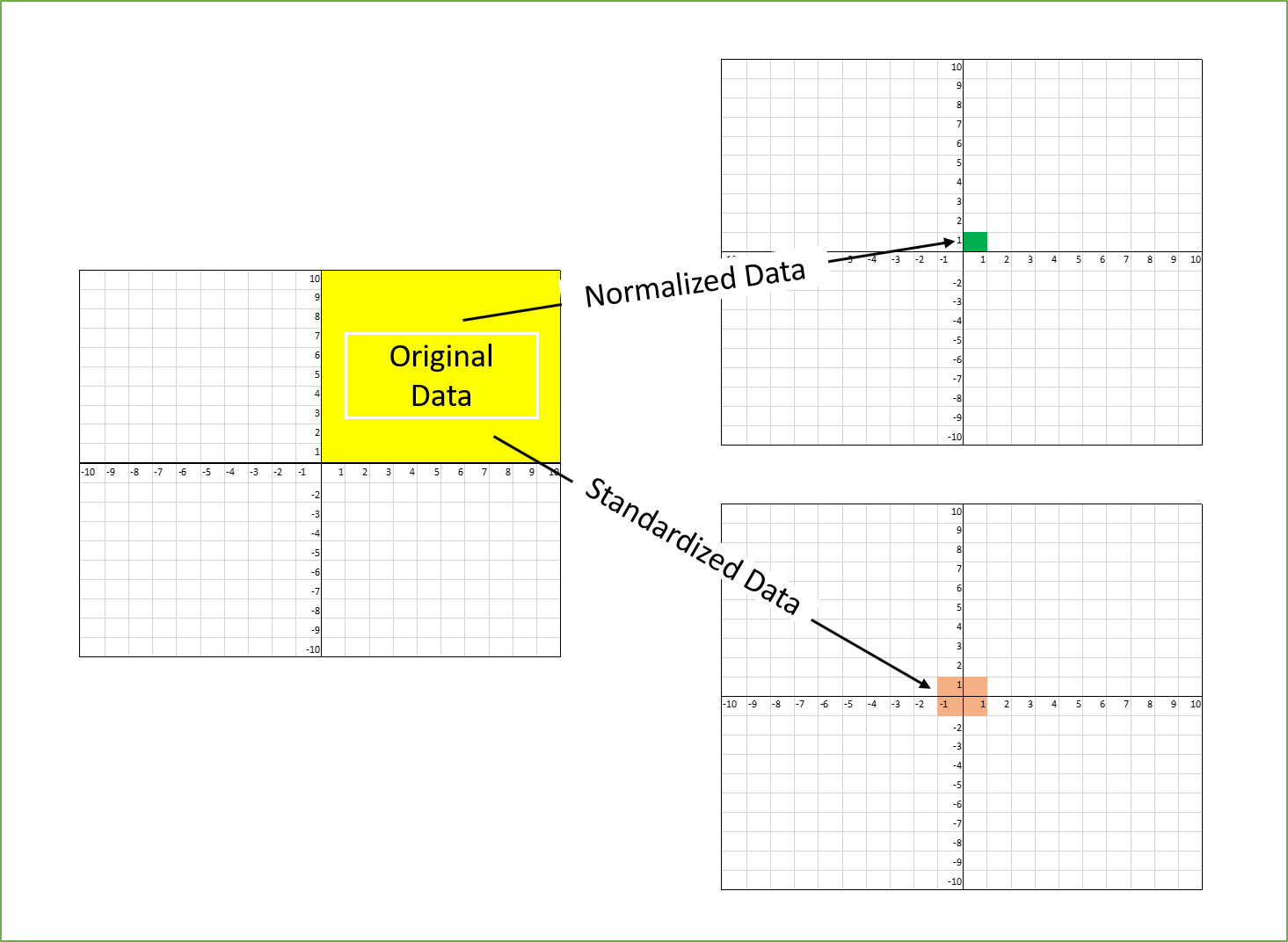
Pearson correlation is used in thousands of real-life situations. For example, scientists in China wanted to know if there was a relationship between how weedy rice populations are different genetically. The goal was to find out the evolutionary potential of the rice. Pearson’s correlation between the two groups was analysed. It showed a positive Pearson Product Moment correlation of between 0.783 and 0.895 for weedy rice populations. This figure is quite high, which suggested a strong relationship.

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

Ans. Scaling in machine learning is one of the most critical steps during the pre-processing of data before creating a machine learning model. Scaling can make a difference between a weak machine learning model and a better one.

The most common techniques of feature scaling are Normalization and Standardization.

Normalization is used when we want to bound our values between two numbers, typically, between [0,1] or [-1,1]. While Standardization transforms the data to have zero mean and a variance of 1, they make our data unitless. Refer to the below diagram, which shows how data looks after scaling in the X-Y plane.



Machine learning algorithm just sees number — if there is a vast difference in the range say few ranging in thousands and few ranging in the tens, and it makes the underlying assumption that higher ranging numbers have superiority of some sort. So, these more significant number starts playing a more decisive role while training the model.

The terms normalization and standardization are sometimes used interchangeably, but they usually refer to different things. Normalization usually means to scale a variable to have a value between 0 and 1, while standardization transforms data to have a mean of zero and a standard deviation of 1.

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

Ans. When R2 reaches 1, VIF reaches infinity.

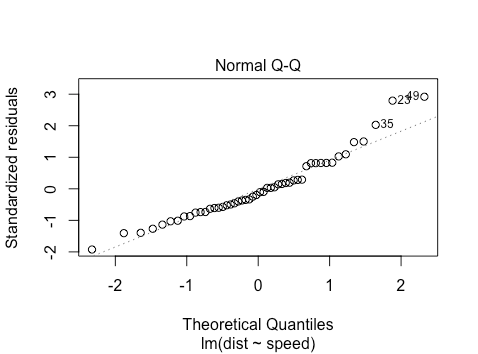
1. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**

Ans. Quantile-Quantile (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.

fit OLS on an R dataset and then analyse the resulting QQ plots.

1. Model<- lm(dist~speed,data=cars)
2. Plot(model)

The second plot will look as follows



The points approximately fall on the line, but what does this mean? The simplest explanation is as follows: say you have some observations and you want to check if they come from a normal distribution. You can standardize them (mean center and scale variance to 1

) and then ‘percentile match’ against a standard normal distribution. Then you can plot your points against a perfectly percentile-matched line.

In more detail, on the x-axis are the theoretical quantiles of a standard normal. That is, we sort the n

points, and then for each i, using the standard normal quantile function we find the x so that P_{\textrm{std norm}}(X\leq x)=\frac{i-0.5}{n}. For this dataset, for the case of the leftmost point, we have that i=1 and n=50. Thus

|  |  |
| --- | --- |
| 1  2 | qnorm(0.5/50)   [1] -2.326348 |

which looks similar to where the leftmost point is on the x-axis. Intuitively, what this is saying is: we have 50 points and we want their x-values to be such that P_{\textrm{std norm}}(X\leq x)=0.01,0.03,\cdots,0.99

. Based on the standard normal distribution, what x do we need to choose? For the y-axis, consider the empirical distribution function of the standardized residuals. We want our corresponding y to be P_{\textrm{emp}}(Y\leq y)=0.01,0.03,\cdots,0.99, but based on the empirical CDF of the standardized residuals.

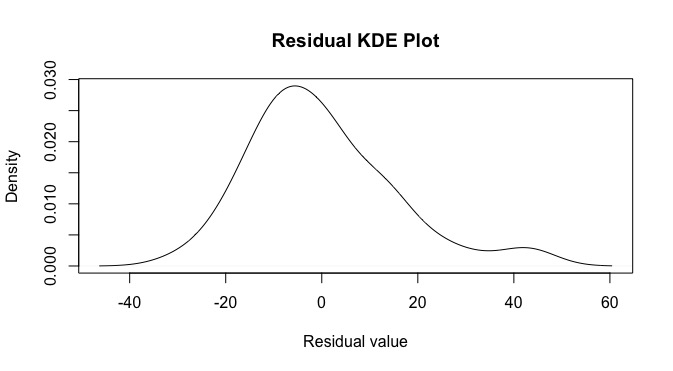
Now how can we characterize the (slight) non-normality? What we see is that on the right-hand side of the graph, the points lie slightly above the line. For the very right-most point, this is saying that the value x

such that P(X < = x) = 0.99 is larger under the empirical CDF for the standardized residuals than it is under a normal distribution. This suggests a ‘fat tail’ on the right-hand side of the distribution.

Other Checks for Normality

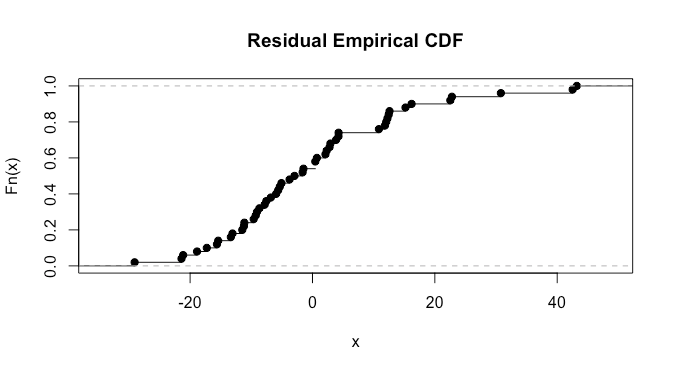
We can investigate further in three ways: a density plot, an empirical CDF plot, and a normality test. Note that one should generally do the former two after the qq plot, as it’s easiest to see that there are departures from normality in a qq plot, but it is sometimes easier to characterize them in density or empirical CDF plots. We can make a density plot as follows:

|  |  |
| --- | --- |
| 1  2 | d<-density(model[['residuals']])  plot (d,main='Residual KDE Plot',xlab='Residual value') |



looking here, there appears to be some negative skewness along with a fat tail on the right. We can also make an empirical CDF plot.

|  |  |
| --- | --- |
| 1 | plot(ecdf(model[['residuals']])) |



here it’s a bit difficult to say much. We can also do a normality test. I ‘suspect’ from looking at the plots that we’ll have a p-value in the .025-0.075 range, as there are clearly some violations, but they don’t look extremely severe: let’s try the Shapiro test.

|  |  |
| --- | --- |
| 1  2  3  4  5  6 | shapiro.test(model[['residuals']])        Shapiro-Wilk normality test    data:  model[['residuals']]  W = 0.94509, p-value = 0.02152 |

The p-value is slightly outside the range I guessed but not by much.