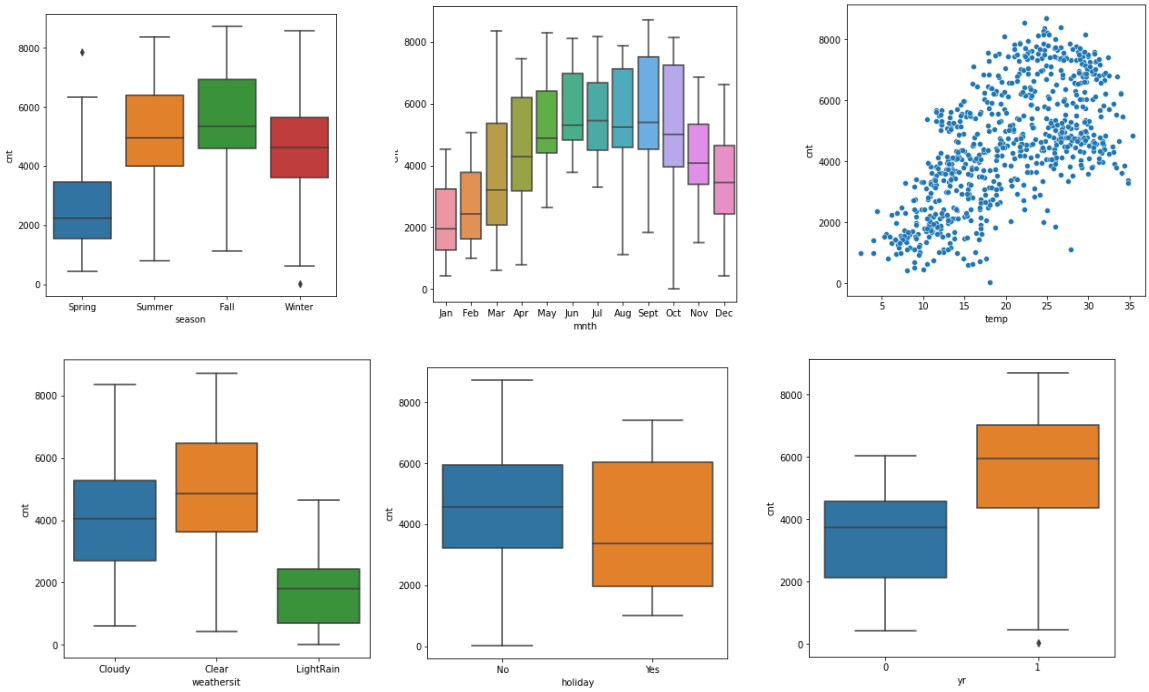
Bike Sharing Assignment Subjective Questions

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.
    a. Bike demand in the fall (Jun, Jul, Aug, Sep) is the highest.
    b. Bike demand is high in the months from May to October(mid-summer to early winter).
    c. Bike demand takes a dip in spring (Dec, Jan, Feb, Mar).
    d. Bike demand is increasing as the temp increases.
    e. Bike demand in year 2019 is higher as compared to 2018.
    f. Bike demand is high if weather is clear or with mist cloudy while it is low when there is light rain or light snow.
    g. B Bike demand is lower when there is a holiday.
    h. Bike demand is almost similar throughout the weekdays and bike demand doesn't change whether day is working day or not.



---

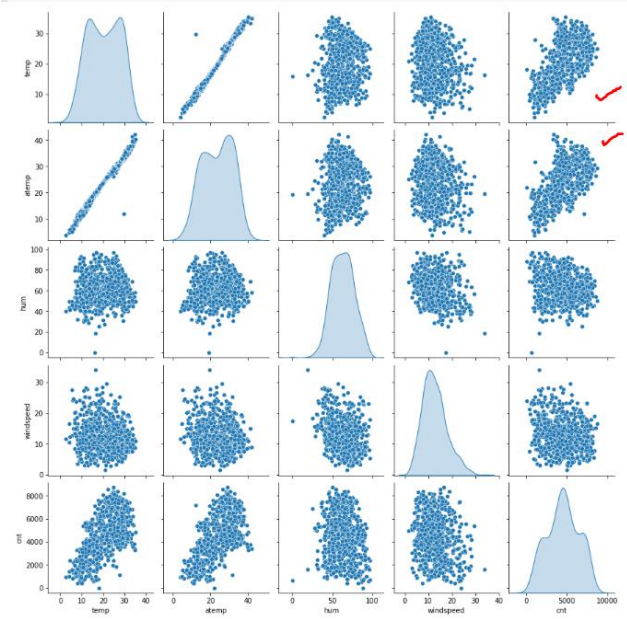
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

    This is done to delete extra column while creating dummy variables. Deleting the column helps achieve k-1 dummy variables to reduce the collinearity between dummy variables. The analysis treats the missing dummy variable as a baseline with which to compare all others.

Example: Consider Weathersit column in the assignment. We have three variables for this: Clear, Cloudy and Light Rain. We can take only 2 variable – 01 and 10 and the base will be 00 which can be removed. 00 indicates Clear, 01 will indicates Cloudy and 10 will indicate it is Light Rain.

---

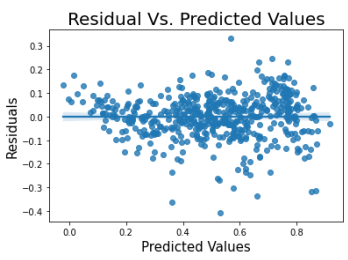
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   Temp and atemp are highly correlated (62% and 63%) with the target variable.

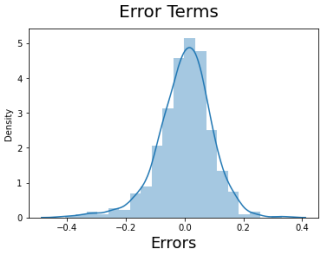


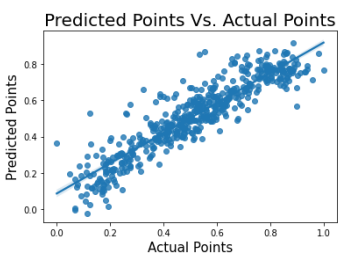
---

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

   To validate the assumptions of linear regression model –







   The plots confirm the below assumptions –

   1. Linear relationship exists between X and y. In the residual-versus-predicted-plot, the points are symmetrically distributed around horizontal line with a roughly constant variance.
   2. Error term is normally distributed with mean 0.
   3. Error terms are independent of each other.
   4. Error terms have constant variance. The data is said to homoscedastic when the residuals are equal across the line of regression. In other words, the variance is equal. There is no visible patterns in the residuals.

5. No Multicollinearity in the data. Multicollinearity was checked using VIF – All VIF are below 5 which is a good indicator that these don't have to be eliminated.

	Features	VIF
1	temp	3.6800
2	windspeed	3.0500
0	yr	2.0000
3	season_Summer	1.5600
7	weathersit_Cloudy	1.4800
4	season_Winter	1.3800
5	mnth_Sept	1.2000
8	weathersit_LightRain	1.0800
6	holiday_Yes	1.0300

---

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

Top 3 features are -

- Temp – With temperature increases, demand for bike is increasing. 0.5480 is the co-efficient of temp which means for every unit increase in temp there is 0.54 unit increase in demand.
- Weather – Light Snow/ Rain is causing decline in the demand. -0.2838 is the co-efficient of temp which means for every unit increase in weather there is -0.2838 unit decline in demand
- Year - Bike demand is growing over the year. 0.2328 is the co-efficient of temp which means for every unit increase in year there is 0.2328 unit increase in demand.

General Subjective Questions

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

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behavior of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Where a and b given by the formulas:

$$b(slobe) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

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

---

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

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

Simple understanding:

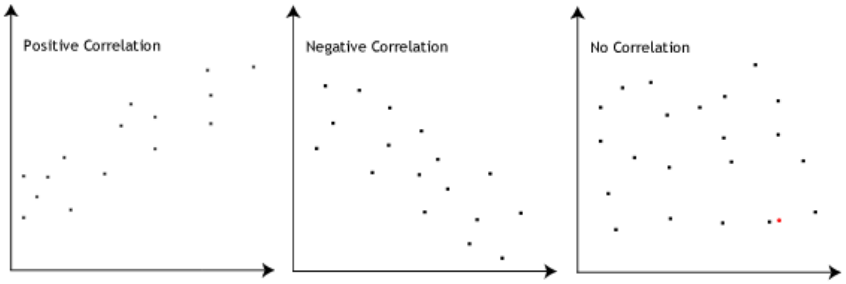
Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

---

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

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.



The Pearson's correlation coefficient varies between -1 and +1 where:

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
r = 0 means there is no linear association
r > 0 < 5 means there is a weak association
r > 5 < 8 means there is a moderate association
r > 8 means there is a strong association

Pearson r Formula

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

Here,

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

---

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

Scaling is a method to make sure that all data is internally consistent. It is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation. So it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale.

- Standardisation basically brings all of the data into a standard normal distribution with mean zero and standard deviation one.
- MinMax scaling, on the other hand, brings all of the data in the range of 0 and 1.

The formulae in the background used for each of these methods are as given below:

- Standardisation: $x = \frac{x - mean(x)}{sd(x)}$
- MinMax Scaling: $x = \frac{x - min(x)}{max(x) - min(x)}$

Scaling for categorical variables - Categorical variables cannot used as they are, so they are converted to numeric format.
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.

---

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 = 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 problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

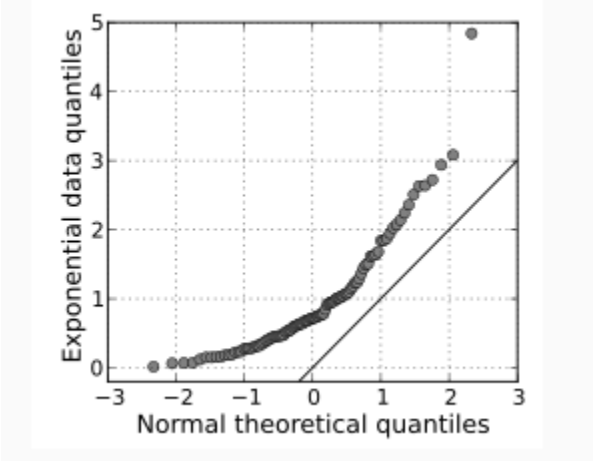
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

---

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

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. 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.

A Q Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.
