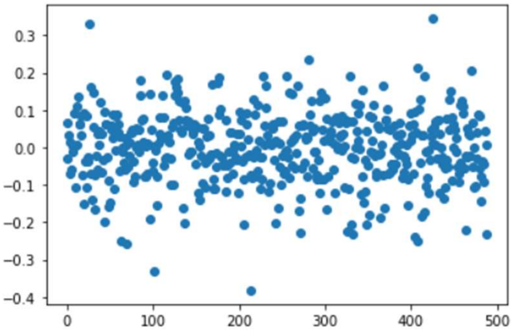
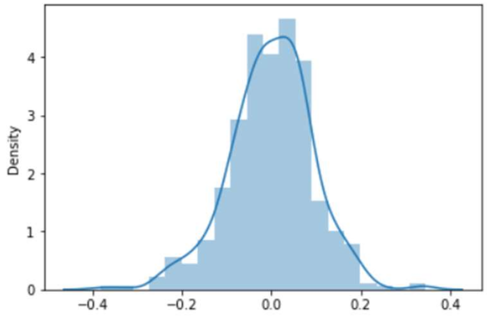
Assignment-based Subjective Questions

1) From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
   - Bike sharing demand:
     i. Warm season and months have higher demands.
     ii. With time, bike-sharing is gaining popularity. Almost 50% increase in YoY basis.
     iii. Bike demands are less during holidays.
     iv. As weather condition deteriorates, bike demand decreases.
2) Why is it important to use **drop_first=True** during dummy variable creation?
   - It's because first category can be represented as negation of rest of the category like:
     i. cat1 => 0000
     ii. cat2 => 1000
     iii. cat3 => 0100
     iv. cat4 => 0010
     v. cat5 => 0001

     Thus, it reduces the correlation among them otherwise first category would be highly corelated with other categories.
3) Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   - temp/atemp (temp & atemp are highly corelated (0.991696)).
4) How did you validate the assumptions of Linear Regression after building the model on the training set?
   - Tried to find pattern in residual using scatter plot.

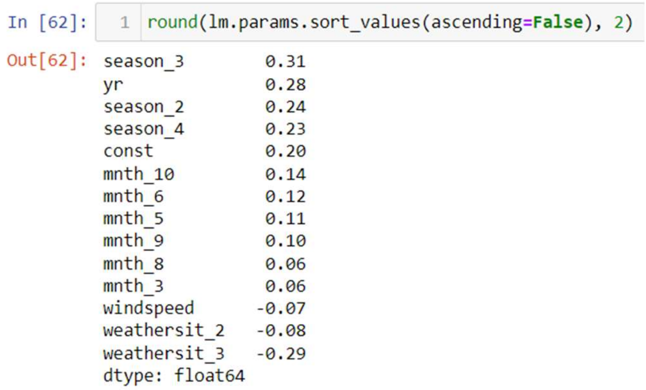


   - Check distribution of residual using distribution plot.



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

- Year, Season & Month



```
In [62]: round(lm.params.sort_values(ascending=False), 2)
```

```
Out[62]: season_3        0.31
         yr              0.28
         season_2        0.24
         season_4        0.23
         const           0.20
         mnth_10         0.14
         mnth_6          0.12
         mnth_5          0.11
         mnth_9          0.10
         mnth_8          0.06
         mnth_3          0.06
         windspeed      -0.07
         weathersit_2   -0.08
         weathersit_3   -0.29
         dtype: float64
```

General Subjective Questions

1) Explain the linear regression algorithm in detail.

Linear Regression is a supervised learning machine algorithm where it tries to predict the target (dependent) variable using independent variable. This is called linear regression because it tries to find a linear relationship of target with independent variables:

$$y_{pred} = \beta_0 + \sum \beta_i * X_i$$

Where i is in set of [1, K].

Here:

β_0 : Intercept on y-axis.

β_i : Coefficient of X_i

Linear regression models tries to update these coefficient in a manner such that the difference between true value and predicted value is minimum.

Root Mean Square Error (RMSE) = $\sum (y_{pred} - y_i)^2 / N$

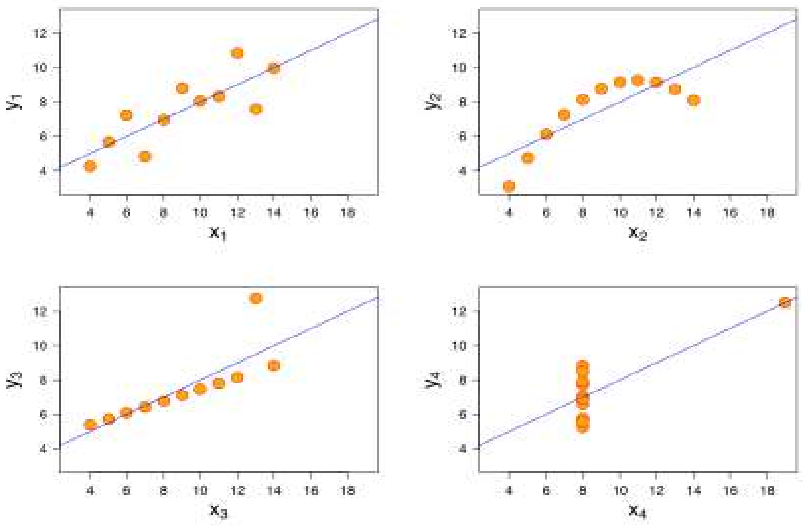
Linear Regression Model (LRM) tries to minimize this RMSE, also known as **cost function** with help of "Gradient Descent Method".

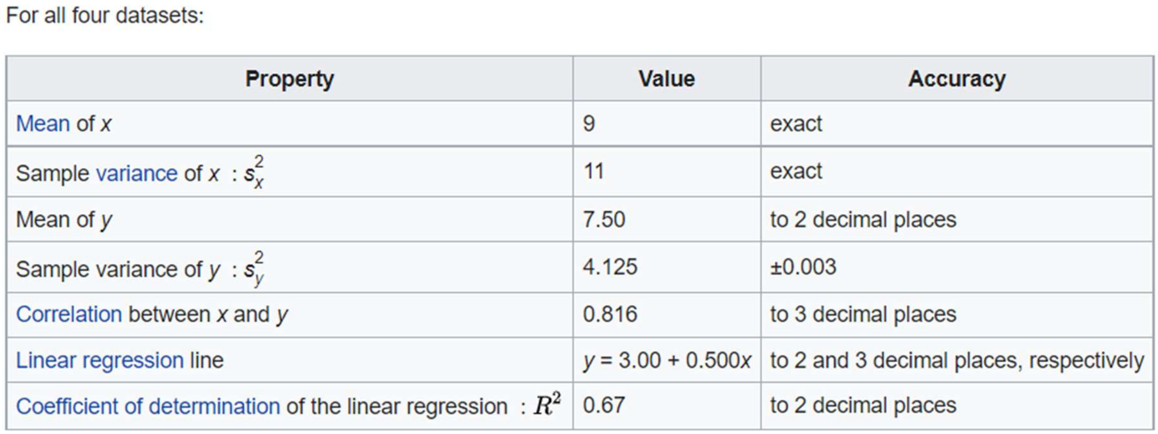
<u>Assumption for LRM:</u>

- There is a linear relationship between independent & dependent variable.
- Error terms ($y_{pred} - y_i$) are normally distributed with mean as zero (0).
- Error terms are independent of each other.
- Error terms have constant variance.

2) Explain the Anscombe’s quartet in detail.

Anscombe’s quartet has four different data sets, having nearly identical simple descriptive statistics, yet very different distribution. Each dataset contains 11 datapoints, and their scatter plot looks very different. It can fool the linear regression model if built.





For all four datasets:

Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x* : s_x^2	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y* : s_y^2	4.125	±0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear regression line	*y* = 3.00 + 0.500*x*	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

(Source: https://en.wikipedia.org/wiki/Anscombe%27s_quartet)

It emphasized the importance of graphical representation of dataset before building a model. In 3rd & 4th dataset, one needs to remove the outliers before building the linear model regression. And 2nd dataset is not a contender of linear regression model as it breaks the assumption of “linear relationship between X & Y”.

3) What is Pearson’s R?

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

Scaling is a method to normalize the range of independent variable. It helps certain machine learning algorithm to train very quickly. If variables are normalized/standardized then “Gradient Descent Method” converge quickly. There are two different methods for scaling:

- Normalized scaling: In this method, values are rescaled in range of 0 to 1. It is also known as Min-Max scaling.

  $X' = (X – X_{min}) / (X_{max} – X_{min})$

- Standard scaling: In this method, values are scaled with respect to mean and standard deviation.

  $X' = (X – \mu) / \sigma$

  Where: σ = standard deviation & μ = mean

In case of standard scaling, values being farther to μ, can have higher value. Thus its range is more. Whereas in normalized scaling, it is ensured that scaled data remains between 0 to 1. Thus, ML methods like “Gradient Descent” and distanced based ML algo like “KNN” “K-means” etc will converge faster in normalized scaling.

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

VIF is a parameter explaining how an independent variable is related with other variable and calculated as

$VIF = 1/(1-R^2)$

If a variable’s variance is completely explained by other variable then its R^2 value would be one, and in that case VIF will be infinite.

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