# 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-**

```
yr              0.234494
workingday      0.100362
temp            0.479076
windspeed      -0.149283
Monday          0.055981
Saturday        0.016295
Sunday          0.114502
September       0.089901
Spring         -0.053930
Summer          0.063278
Winter          0.097467
weathersit_2   -0.081632
weathersit_3   -0.288461
```

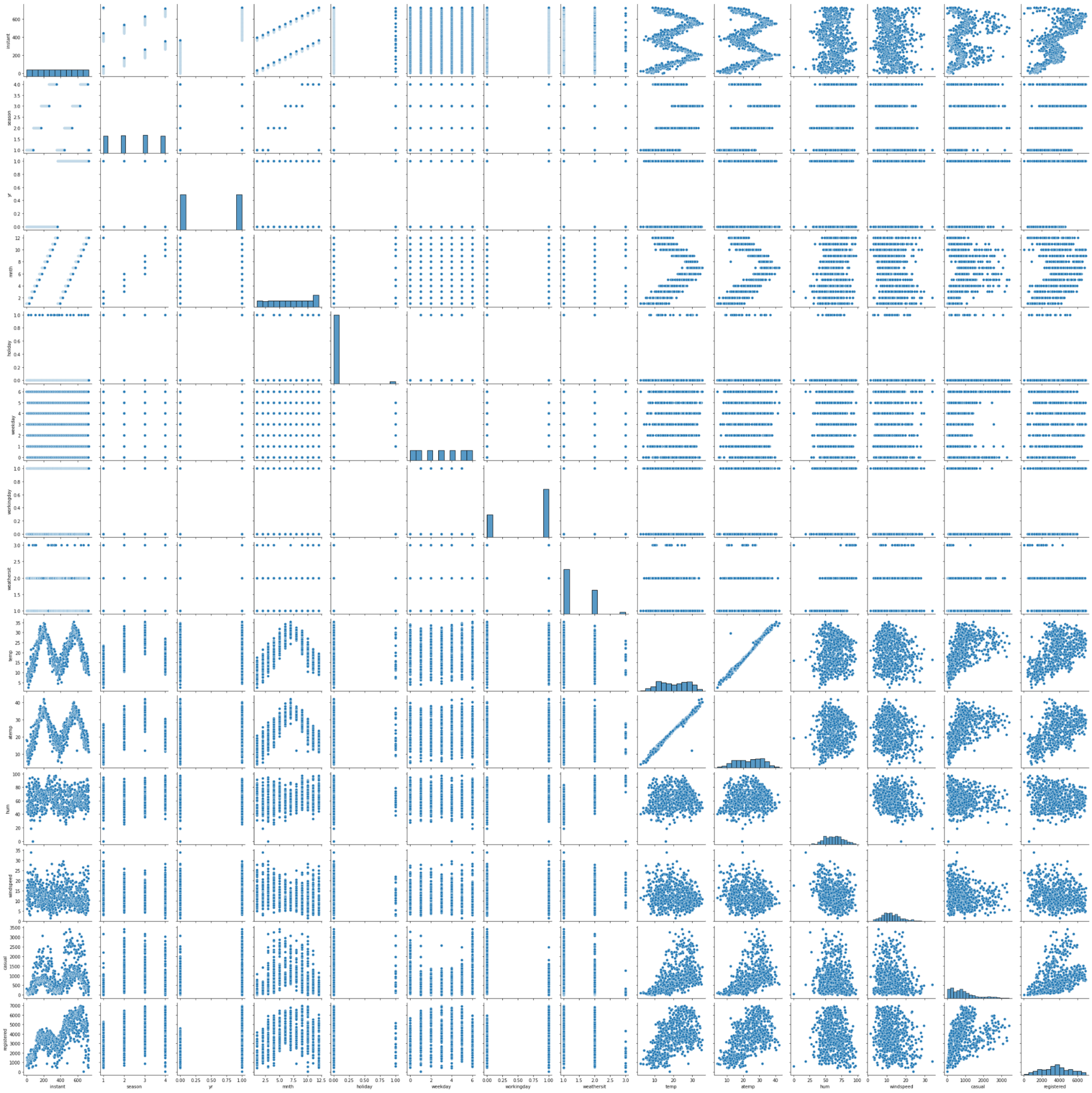
By looking at the above result, we can say **yr**, **workingday**, **temp**, **Monday**, **Saturday**, **Sunday**, **September**,**Summer** and **Winter** affects positively on the bike demands while **weathersit_2**, **weathersit_3**, **windspeed** and **Spring** negatively affects the bike demands.

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

**Ans-** It is very important to use **drop_first=True** during the dummy variable creation because **it helps in reducing the extra column created during dummy variable creation**. Hence it reduces the correlations created among dummy variables.

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

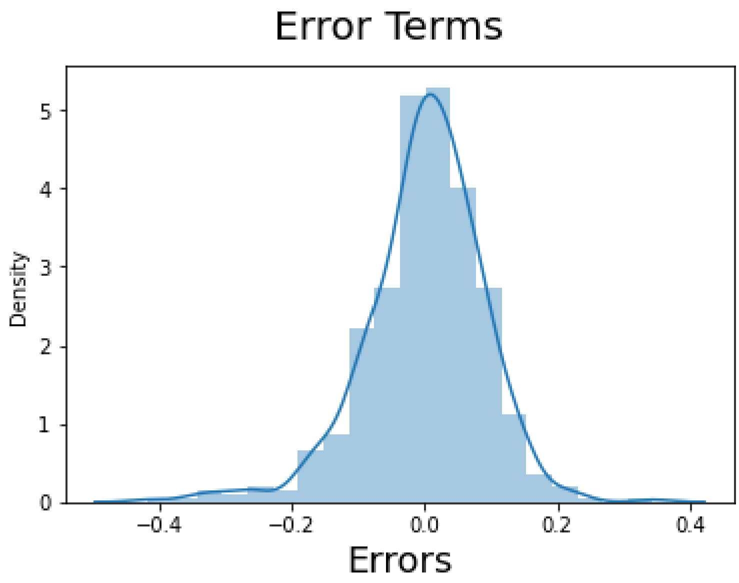
**Ans-**



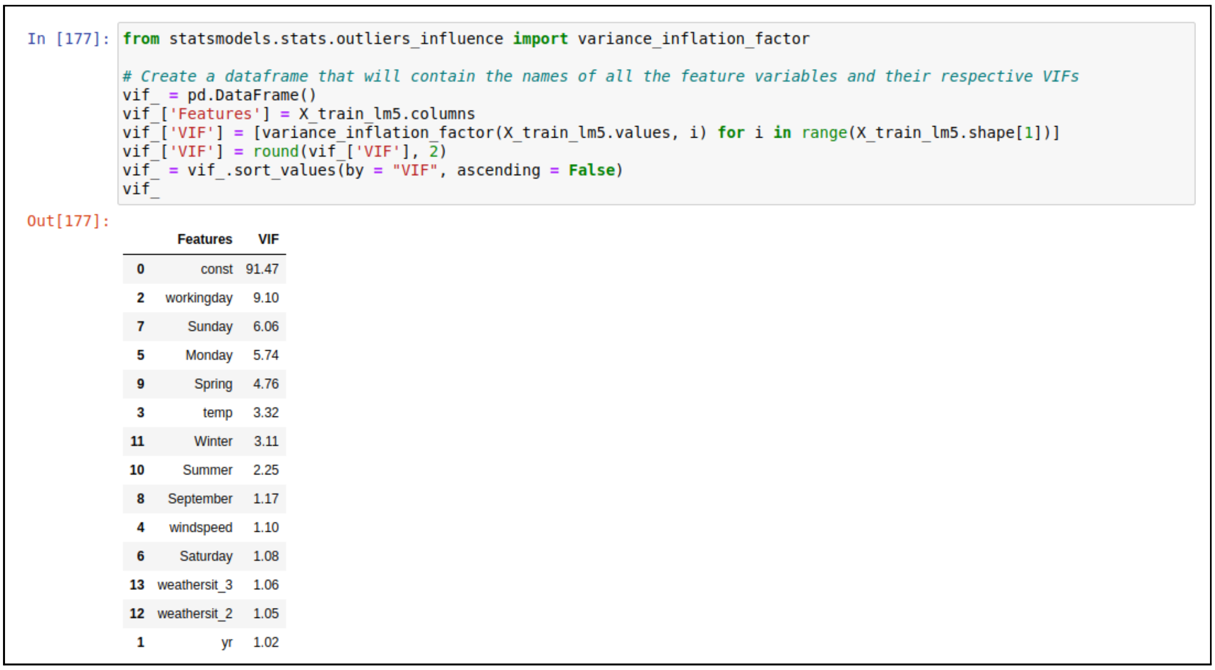
After looking at the above pairplot among the numerical columns, it showed that the column **temp** and **atemp** are highly correlated to the target variable **cnt**.

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

**Ans-** For validating the assumptions of lInear regression after building the model on the training set, I plotted the distribution plot for training errors and saw that the errors are distributed normally which validated the first assumptions of linear regression. The plot has been shown below:



For validating the second assumption which is checking for multicollinearity in variables, I calculated the VIF (variance inflation factor) and I found that 2 features were having VIF value more than 5 but less than 10. So considering the threshold of dropping the feature if correlated to be 10, we can say our second assumption for linear regression passed too. Below is the table I received.



```
In [177]: from statsmodels.stats.outliers_influence import variance_inflation_factor

          # Create a dataframe that will contain the names of all the feature variables and their respective VIFs
          vif_ = pd.DataFrame()
          vif_['Features'] = X_train_lm5.columns
          vif_['VIF'] = [variance_inflation_factor(X_train_lm5.values, i) for i in range(X_train_lm5.shape[1])]
          vif_['VIF'] = round(vif_['VIF'], 2)
          vif_ = vif_.sort_values(by = "VIF", ascending = False)
          vif_
Out[177]:
```

| | Features | VIF |
|---|---|---|
| 0 | const | 91.47 |
| 2 | workingday | 9.10 |
| 7 | Sunday | 6.06 |
| 5 | Monday | 5.74 |
| 9 | Spring | 4.76 |
| 3 | temp | 3.32 |
| 11 | Winter | 3.11 |
| 10 | Summer | 2.25 |
| 8 | September | 1.17 |
| 4 | windspeed | 1.10 |
| 6 | Saturday | 1.08 |
| 13 | weathersit_3 | 1.06 |
| 12 | weathersit_2 | 1.05 |
| 1 | yr | 1.02 |

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

**Ans-** As per our final Model, the top 3 predictor variables that influences the bike booking are:

- **Temperature (temp)** - A coefficient value of ‘`0.479076`’ indicated that a unit increase in temp variable increases the bike hire numbers by `0.479076` units.

- **Weather Situation 3 (weathersit_3)** - A coefficient value of '-`0.288461`' indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by `0.288461` units.
- **Year (yr)** - A coefficient value of '`0.234494`' indicated that a unit increase in yr variable increases the bike hire numbers by `0.234494` units.

# General Subjective Questions

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

**Ans-** Linear regression is the method of finding the best straight line that can fit the given set of data, but this can work only when there is a linear relationship between the dependent (target variables) and independent variables (feature variables). Linear regression finds the relationship between independent and dependent variables.

There are few assumptions that the linear regression takes into consideration, they are mentioned below:

1. It is assumed that there is a linear relationship between the dependent and independent variables. It is known as the 'linearity assumption'.
2. Assumptions about the residuals:
    1. Normality assumption: It is assumed that the error terms, $\varepsilon^{(i)}$, are normally distributed.
    2. Zero mean assumption: It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.
    3. Constant variance assumption: It is assumed that the residual terms have the same (but unknown) variance, $\sigma^2$ . This assumption is also known as the assumption of homogeneity or homoscedasticity.
    4. Independent error assumption: It is assumed that the residual terms are independent of each other, i.e. their pairwise covariance is zero.
3. Assumptions about the estimators:
    1. The independent variables are measured without error.
    2. The independent variables are linearly independent of each other, i.e. there is no multicollinearity in the data.

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

**Ans-** 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. It is mentioned in the definition that Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistical properties for describing realistic datasets.

**3. What is Pearson’s R?**
**Ans-** In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

There are certain requirements for Pearson's Correlation Coefficient:
- Scale of measurement should be interval or ratio
- Variables should be approximately normally distributed
- The association should be linear
- There should be no outliers in the data

Formula is given as:

$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2-(\Sigma x)^2][N\Sigma y^2-(\Sigma y)^2]}}$$

Where,

**N =** the number of pairs of scores

**Σxy =** the sum of the products of paired scores

**Σx =** the sum of x scores

**Σy =** the sum of y scores

**Σx2 =** the sum of squared x scores

**Σy2 =** the sum of squared y scores

**4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
**Ans-** It is a step of data Preprocessing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in

an algorithm. Most of the time, the collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then the algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

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.

**Normalized Scaling (MinMax scaling)**

- It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

MinMax scaling x =  (x - min(x) ) / (max(x) - min(x))

**Standardized Scaling**

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**).
- One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.

Standardized scaling  x  = x - mean(x) / std deviation (x)

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

**Ans-** An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables. When VIF value shows infinite then it represents that the variable is perfectly correlated hence it becomes the situation of multicollinearity which violates the assumption of Linear regression. In the case of perfect correlation, we get R2 =1, which leads 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.

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

**Ans-** 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 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.

**Submitted By: Shruti Sneha**