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? (3 marks)**

   Below are the inferences

   a. **Working Days** – Demand is high during working days
   b. **Season** – Demand is high during Summer season. Demand is less during Winter and spring season
   c. **Weather Situation** – Demand is high when weather is clear and and less during Mist Cloudy and Light Snow
   d. **Year** – Demand was more in 2019 compared to 2018
   e. **Temperature** - Demand increases with temperature
   f. **Windspeed** - Demand decreases with windspeed

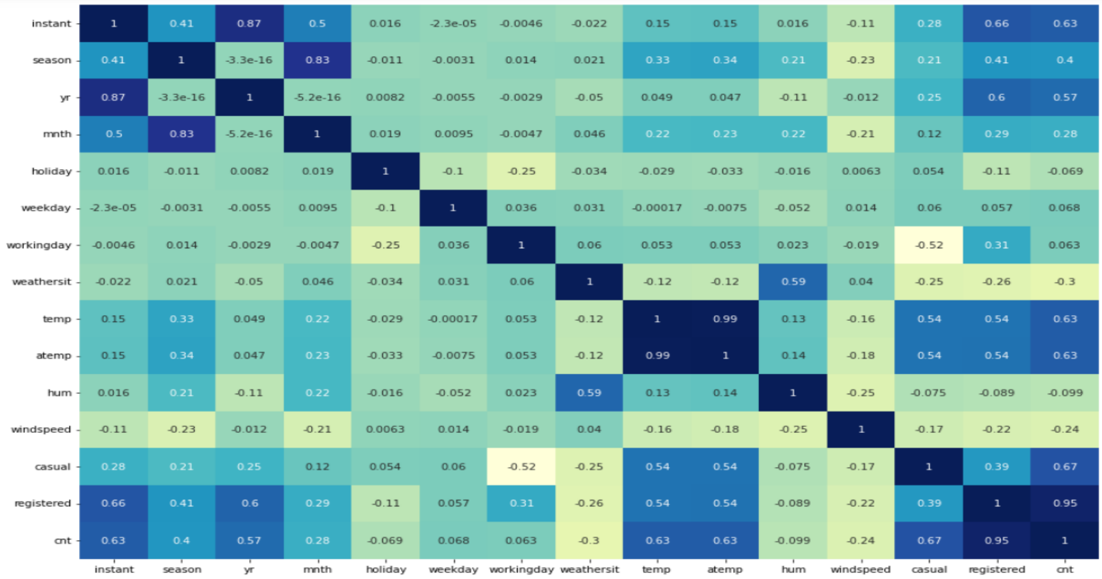
2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**

   This is to avoid redundancy. The importance of this variable is actually contributed by other variables and it helps with resolving Multi-Colinearity in Multiple Linear Regression

   Keeping k dummies for k levels of a categorical variable is not needed as there is a redundancy of one level, which is here in separate column. This is not needed since one of the combinations will be uniquely representing this redundant column. Hence, it's better to drop one of the columns and just have k-1 dummies(columns) to represent k levels.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

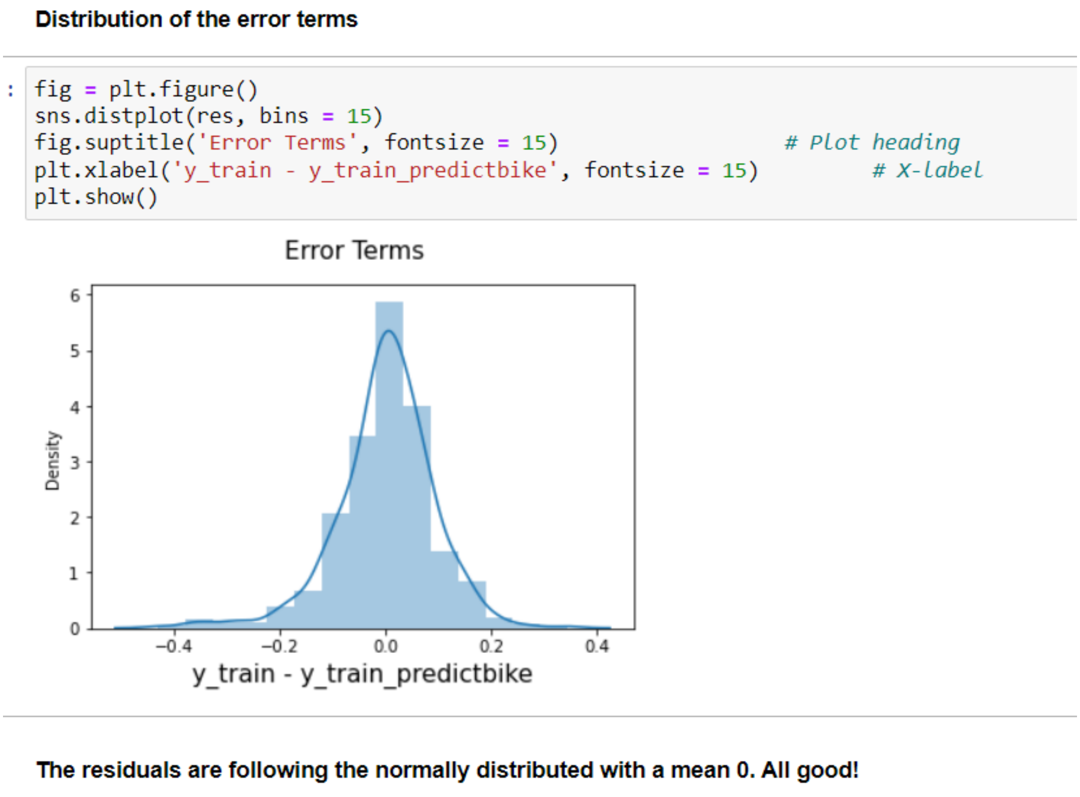
   Correlation is highest between **temp and atemp (0.99)**



	instant	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
instant	1	0.41	0.87	0.5	0.016	-2.3e-05	-0.0046	-0.022	0.15	0.15	0.016	-0.11	0.28	0.66	0.63
season	0.41	1	-3.3e-16	0.83	-0.011	-0.0031	0.014	0.021	0.33	0.34	0.21	-0.23	0.21	0.41	0.4
yr	0.87	-3.3e-16	1	-5.2e-16	0.0082	-0.0055	-0.0029	-0.05	0.049	0.047	-0.11	-0.012	0.25	0.6	0.57
mnth	0.5	0.83	-5.2e-16	1	0.019	0.0095	-0.0047	0.046	0.22	0.23	0.22	-0.21	0.12	0.29	0.28
holiday	0.016	-0.011	0.0082	0.019	1	-0.1	-0.25	-0.034	-0.029	-0.033	-0.016	0.0063	0.054	-0.11	-0.069
weekday	-2.3e-05	-0.0031	-0.0055	0.0095	-0.1	1	0.036	0.031	-0.00017	-0.0075	-0.052	0.014	0.06	0.057	0.068
workingday	-0.0046	0.014	-0.0029	-0.0047	-0.25	0.036	1	0.06	0.053	0.053	0.023	-0.019	-0.52	0.31	0.063
weathersit	-0.022	0.021	-0.05	0.046	-0.034	0.031	0.06	1	-0.12	-0.12	0.59	0.04	-0.25	-0.26	-0.3
temp	0.15	0.33	0.049	0.22	-0.029	-0.00017	0.053	-0.12	1	0.99	0.13	-0.16	0.54	0.54	0.63
atemp	0.15	0.34	0.047	0.23	-0.033	-0.0075	0.053	-0.12	0.99	1	0.14	-0.18	0.54	0.54	0.63
hum	0.016	0.21	-0.11	0.22	-0.016	-0.052	0.023	0.59	0.13	0.14	1	-0.25	-0.075	-0.089	-0.099
windspeed	-0.11	-0.23	-0.012	-0.21	0.0063	0.014	-0.019	0.04	-0.16	-0.18	-0.25	1	-0.17	-0.22	-0.24
casual	0.28	0.21	0.25	0.12	0.054	0.06	-0.52	-0.25	0.54	0.54	-0.075	-0.17	1	0.39	0.67
registered	0.66	0.41	0.6	0.29	-0.11	0.057	0.31	-0.26	0.54	0.54	-0.089	-0.22	0.39	1	0.95
cnt	0.63	0.4	0.57	0.28	-0.069	0.068	0.063	-0.3	0.63	0.63	-0.099	-0.24	0.67	0.95	1

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

a) Distribution of error terms

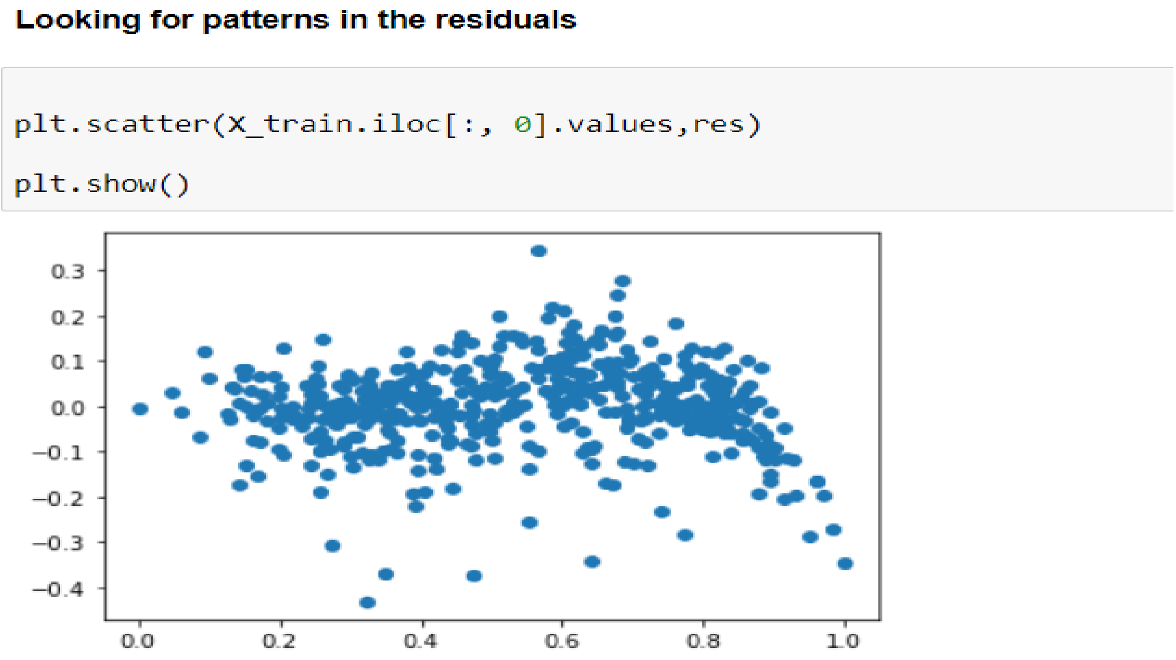


Distribution of the error terms

```
fig = plt.figure()
sns.distplot(res, bins = 15)
fig.suptitle('Error Terms', fontsize = 15)                  # Plot heading
plt.xlabel('y_train - y_train_predictbike', fontsize = 15)         # X-label
plt.show()
```

The residuals are following the normally distributed with a mean 0. All good!

b) Looking for patterns



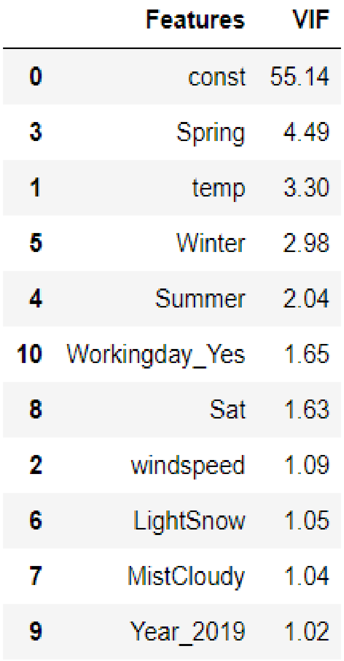
Looking for patterns in the residuals

```
plt.scatter(X_train.iloc[:, 0].values,res)

plt.show()
```

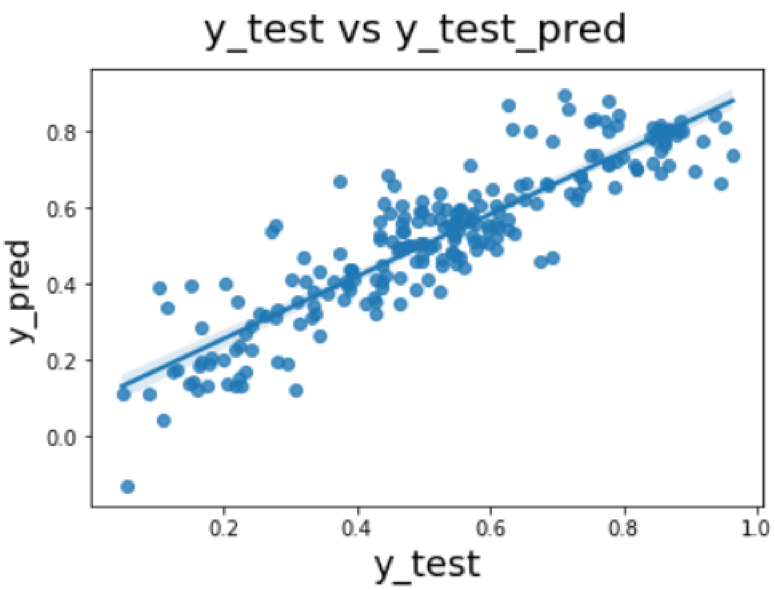
c) No Multi-collinearity

Final VIFs are below 5 for all Features



	Features	VIF
0	const	55.14
3	Spring	4.49
1	temp	3.30
5	Winter	2.98
4	Summer	2.04
10	Workingday_Yes	1.65
8	Sat	1.63
2	windspeed	1.09
6	LightSnow	1.05
7	MistCloudy	1.04
9	Year_2019	1.02

d) Homoscedastic



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

# list of coefs based on final model

```
temp            0.468757
windspeed       -0.156150
Spring         -0.081129
Summer          0.038572
Winter          0.078210
LightSnow       -0.284259
MistCloudy      -0.078237
Sat            0.066801
Year_2019       0.234246
Workingday_Yes   0.055724
```

Top 3 features contributing significantly towards explaining the demand are

1. temp,
2. year and
3. Season

General Subjective Questions

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

   Linear Regression is a machine learning algorithm based on **supervised learning**. Linear Regression models a target value based on independent variables. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variables (x1, x2, …).

   There are two types of linear regression

   a) **Simple Linear Regression** The most elementary type of regression model is the simple linear regression which explains the relationship between a dependent variable (y) and one independent variable (x) using a straight-line. The straight line is plotted on the scatter plot of these two points.

      This is represented by simple equation
      y = mx + b
      m=slope/ coefficient and b = intercept

   b) **Multiple linear regression** is a technique to understand the relationship between one dependent variable (y) and several independent variables. T**he** objective of multiple regression is to find a linear equation that can best determine the value of dependent variable y for different values independent variables (x1,x2,x3…xn).

      This is represented by equation

      Y = m1x1+m2x2+ ….. + mnxn + b

      M1,m2, mn = coefficients and b = intercept

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

   Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

a) **Dataset 1:** this fits the linear regression model pretty well.

b) **Dataset 2:** this could not fit linear regression model on the data quite well as the data is non-linear.

c) **Dataset 3:** shows the outliers involved in the dataset which cannot be handled by linear regression model

d) **Dataset 4:** shows the outliers involved in the dataset which cannot be handled by linear regression model

It was constructed by statistician Francis Anscombe to illustrate
the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties. This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets

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

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Pearson correlation coefficient (r)	**Correlation type**	**Direction**
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the same direction.
0	No correlation	There is no relationship between the variables.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the opposite direction.

The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.

Pearson correlation coefficient (r)	**Correlation type**	**Strength**
Greater than 0.5	Positive correlation	Strong
Between 0.3 and 0.5	Positive correlation	Moderate
Between 0 and .3	Positive correlation	Weak
0	None	None
Between 0 and - .3	Negative correlation	Weak
Between - 0.3 and - 0.5	Negative correlation	Moderate
Greater than - 0.5	Negative correlation	Strong

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

Scaling is a technique to standardize the independent features present in the data in a fixed range.

When we have many variables in a model, a lot of them might be on very different scales which will lead a model with coefficients that might be difficult to interpret. So, we need to scale features because of two reasons:

1. Ease of interpretation

2. Faster convergence for gradient descent methods

We can scale the features using two very popular method

1. Min-Max scaling or Normalized Scaling
2. Standardized Scaling

Normalized Scaling	**Standardized Scaling**
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
Scales values between [0, 1] or [-1, 1]	It is not bounded to a certain range
It is really affected by outliers.	It is much less affected by outliers
Scikit-Learn provides a transformer called MinMaxScaler for Normalization	Scikit-Learn provides a transformer called StandardScaler for standardization.

It is useful when we don’t know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
It is used when features are of different scales.	It is useful when the feature distribution is Normal or Gaussian.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.

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).

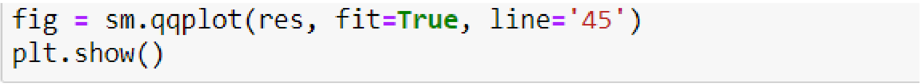
6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

   Q-Q plot is a graphical plotting of the quantiles of two distributions with respect to each other. In other words we can say plot quantiles against quantiles. Whenever we are interpreting a Q-Q plot, we shall concentrate on the ‘y = x’ line. We also call it the 45-degree line in statistics. It entails that each of our distributions has the same quantiles. In case if we witness a deviation from this line, one of the distributions could be skewed when compared to the other.

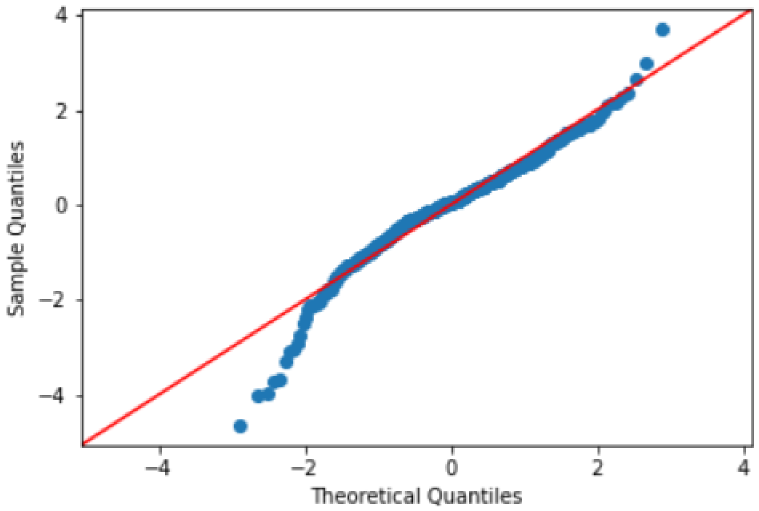
   If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical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

Q-Q plot from Assignment model



```
fig = sm.qqplot(res, fit=True, line='45')
plt.show()
```


Sample Quantiles
Theoretical Quantiles