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)

Answer 1:

- Almost 97% of the booking we made in holiday.
- Almost 68% bookings were made during the clear weather followed by almost 30% in Misty weather
- Weekday shows close trend, every weekday contributes around 13% to 14%
- Overall the bike rental demand has gone up from 2018 to 2019

---

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

Answer 2:

- drop_first=True is important to use, as it helps in reducing the redundant column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
- Redundant column in the context, can be inferred by other columns. For example column season, there are 4 season, so if 3 columns represent 3 season, if there is no occurrence of 3 season that infers that 4^{th} season occur

---

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

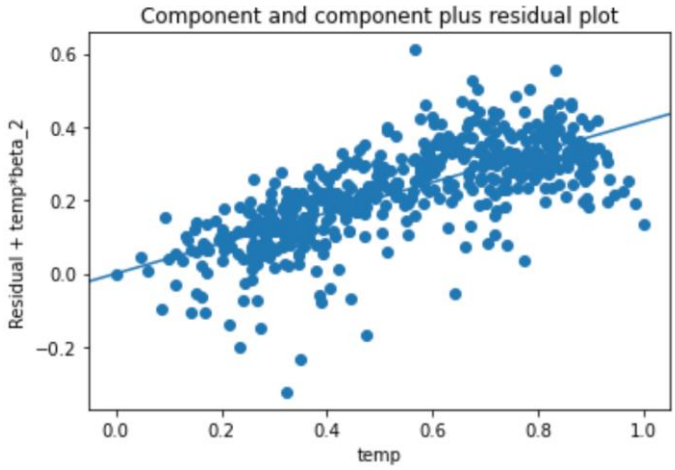
Answer 3:
Cnt vs registered

---

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

1. Linear Relationship
   1.1. linear regression needs the relationship between the independent and dependent variables to be linear. It is also important to check for outliers since linear regression is sensitive to outlier effects. The linearity assumption can best be tested with scatter plots
   1.2. Screenshot from assignment

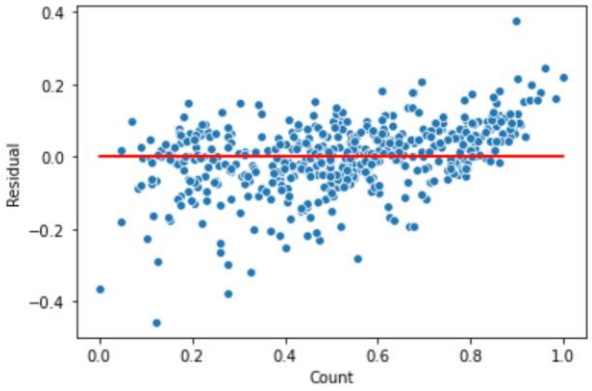
```
sm.graphics.plot_ccpr(lr5, 'temp')
plt.show()
```



2. Homoscedasticity
   2.1. Homoscedasticity means that the residuals have constant variance no matter the level of the dependent variable.
   2.2. Screenshot from assignment

12.2 Homoscedasticity

```
y_train_pred = lr5.predict(X_train_lm5)
residual = y_train - y_train_pred
sns.scatterplot(y_train,residual)
plt.plot(y_train,(y_train - y_train), '-r')
plt.xlabel('Count')
plt.ylabel('Residual')
plt.show()
```

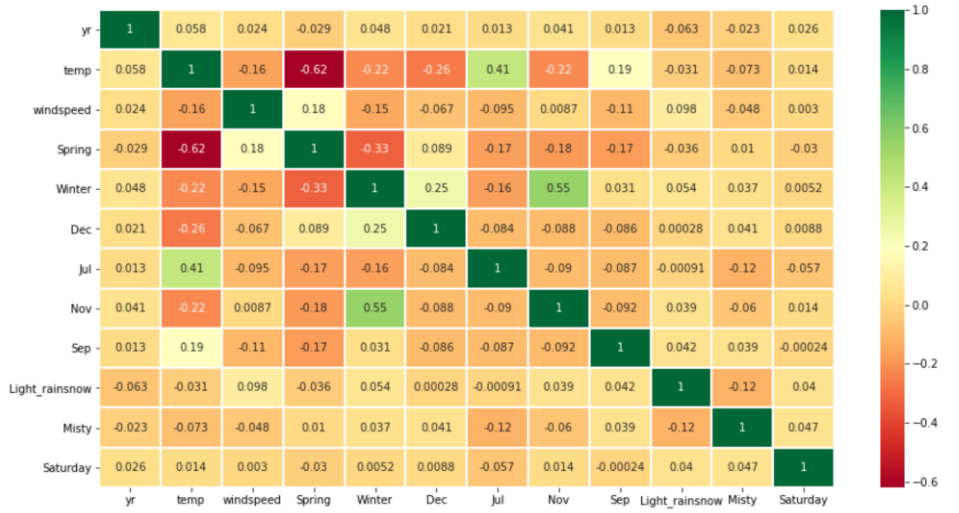


There is no visible pattern in residual values, thus homoscedacity is well preserved

3. Absence of Multicollinearity
   3.1. Multicollinearity refers to the fact that two or more independent variables are highly correlated

3.2. Screenshot from assignment, as we can see there is no high collinearity exist

```
# Validating Multi Colinearity
plt.figure(figsize=(15,8))
sns.heatmap(X_train_new.corr(),annot = True, cmap="RdYlGn",linewidth =1)
plt.show()
```



4. Independence of residuals

4.1. Auto-correlation could mean that the linearity of the relationship is not respected or that variables may have been omitted. Auto-correlation would lead to spurious relationships between the independent variables and the dependent variable.

5. Normality of Errors

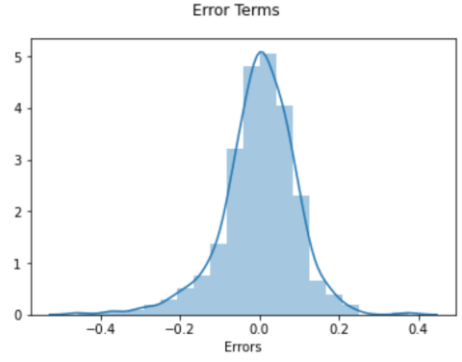
5.1. If the residuals are not normally distributed, Ordinary Least Squares (OLS), and thus the regression, may become biased.

12.5 Normality of error

```
res = y_train-y_train_pred

# Plot the histogram of the error terms
fig = plt.figure()
sns.distplot((res), bins = 20)
fig.suptitle('Error Terms')
plt.xlabel('Errors')
plt.show()
```

5.2.



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

Answer 3:

Temperature, Light_rainsnow and yr are the variables having high impact on bike renting. It is recommended that bike availability and promos to be increased during the summer months.

15. Model Outcome summary

```
1 lr5.summary()
```

OLS Regression Results

Dep. Variable:	cnt	**R-squared:**	0.827
Model:	OLS	**Adj. R-squared:**	0.823
Method:	Least Squares	**F-statistic:**	197.7
Date:	Mon, 10 May 2021	**Prob (F-statistic):**	1.99e-180
Time:	16:55:07	**Log-Likelihood:**	486.59
No. Observations:	510	**AIC:**	-947.2
Df Residuals:	497	**BIC:**	-892.1
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2778	0.026	10.751	0.000	0.227	0.329
yr	0.2333	0.008	27.603	0.000	0.217	0.250
temp	0.4151	0.032	12.865	0.000	0.352	0.479
windspeed	-0.1516	0.027	-5.521	0.000	-0.206	-0.098
Spring	-0.1203	0.016	-7.614	0.000	-0.151	-0.089
Winter	0.0703	0.015	4.831	0.000	0.042	0.099
Dec	-0.0570	0.018	-3.222	0.001	-0.092	-0.022
Jul	-0.0459	0.018	-2.595	0.010	-0.081	-0.011
Nov	-0.0623	0.019	-3.209	0.001	-0.100	-0.024
Sep	0.0537	0.016	3.349	0.001	0.022	0.085
Light_rainsnow	-0.3091	0.027	-11.406	0.000	-0.362	-0.256
Misty	-0.0753	0.009	-8.350	0.000	-0.093	-0.058
Saturday	0.0251	0.012	2.093	0.037	0.002	0.049

General Subjective Questions

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

Answer 1:

Linear regression is used to predict a quantitative response Y from the predictor variable X.

Mathematically, we can write a linear regression equation as:

Where a and b given by the formulas:

$$y = a + bx$$

Where a and b given by the formulas:

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

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

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

Linear regression is use where the output is (Y) continuous

1. The equation of the best fit regression line $Y = \beta_0 + \beta_1 X$ can be found by minimising the cost function (RSS in this case, using the ordinary least squares method), which is done using the following two methods:
    - Differentiation
    - Gradient descent
2. The strength of a linear regression model is mainly explained by R^2, where ****$R^2 = 1 - (RSS/TSS)$.

- **RSS:** Residual sum of squares
- **TSS:** Total sum of squares

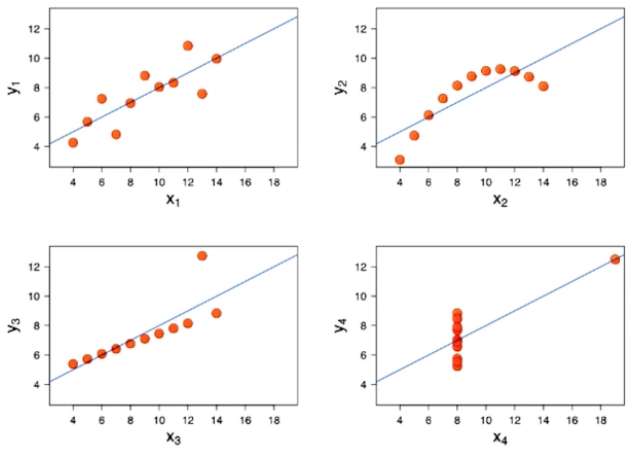
3. RSE helps in measuring the lack of fit of a model on a given data. The closeness of the estimated regression coefficients to the true ones can be estimated using RSE. It is related to RSS by the formula: RSE=√RSSdf, where df=n−2 and n is the number of data points.

---

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

Answer:
Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed
This is used to to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.
For eg: below is the data that has same mean, standard deviation yet they may have linear, non-linear or no relationship.

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



- In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

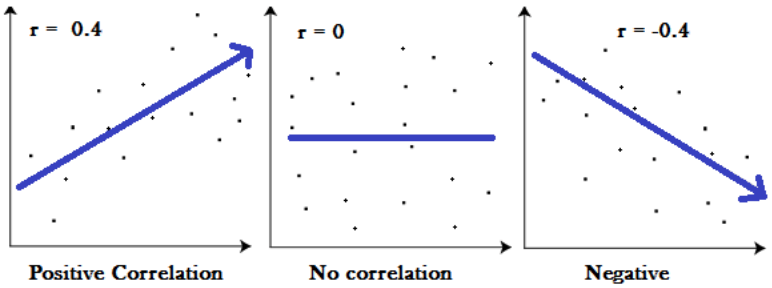
---

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

Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

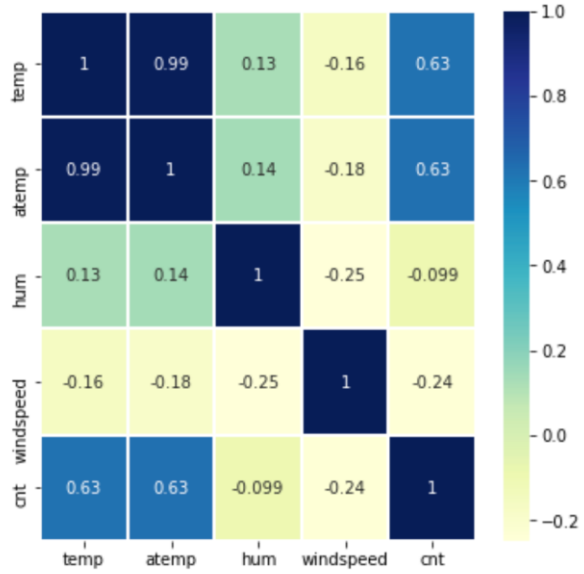
- 1 indicates a strong positive relationship.
- 1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.



Eg: from the assignment

```
# Correlation

plt.figure(figsize = (6,6))
ax= sns.heatmap(bike_num.corr(), annot = True, cmap="YlGnBu",linewidth =1)
```



Mathematical Formula to calculate

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

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

---

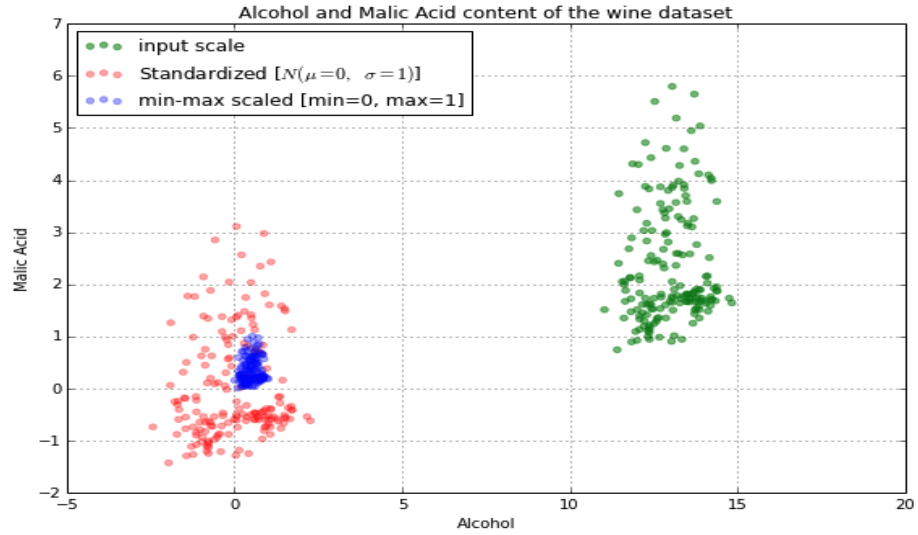
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range.
If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values
There are 2 most important techniques

1. Min-Max Normalization:
    a. This technique re-scales a feature or observation value with distribution value between 0 and 1.
    b. $$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$
2. Standardization:
    a. It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1
    b. $$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

Example as below:

For example:



---

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

If VIF is infinite that means there is a perfect correlation. A large value of VIF indicates that there is a correlation between the variables. That means there is multicollinearity exist. And If there is multicollinearity then it has to be treated like eliminating variable etc.

---

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 also known as Quantile-quantile plot, is a graphical tool to help us asses if a set of data plausibly came from some theoretical distribution such as a Normal or exponential Q-Q plots take your sample data, sort it in ascending order, and then plot them versus quantiles calculated from a theoretical distribution. The number of quantiles is selected to match the size of your sample data

For e.g.: screenshot from assignment

```
1 sm.qqplot((y_train - y_train_pred), fit=True, line='45')
2 plt.show()
```