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) Following the analysis result on categorical variables:

season vs cnt:

- Season 3 which is the fall has maximum number of bookings. (Season 3>2>4>1)

year vs cnt:

- 2019 has a greater number of bookings than previous year which is a good sign of company's growth.

Month vs cnt

- Total number of bookings are more during May-Sep/Oct but then there is a noticeable drop. Same observed during univariate analysis

Week vs cnt

- Total number of bookings mostly same for all the days in week. There is little more demand on Thursday, Friday and Saturday. Little less on Sunday.

weathersit vs cnt:

- '1' which represent clear or partly cloudy weather has more bookings. (Weathersit-1>2>3)

holiday vs cnt:

- During the holidays, there is a dip in total number of bookings, if you compare 25-50 percentile.

workingday vs cnt:

- Workingday or not, they is no noticeable different.

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

A) While converting categorical variables to dummy variable, we need to create n-1 dummy variables to avoid multicollinearity which explains the predictability of an independent variable by other depending variables.

For the bike-lending case study, we have 4 four seasons (1: spring, 2: summer, 3: fall, 4: winter). In this case, 3 dummy variables are created with binary values (2: summer, 3: fall, 4: winter).When all the 3

variables have 0, then it represents 'spring'. In case, all the 4 dummy variables are added to the model, this gives 100% predictability of the 'spring' which effects the interpretability of the model.

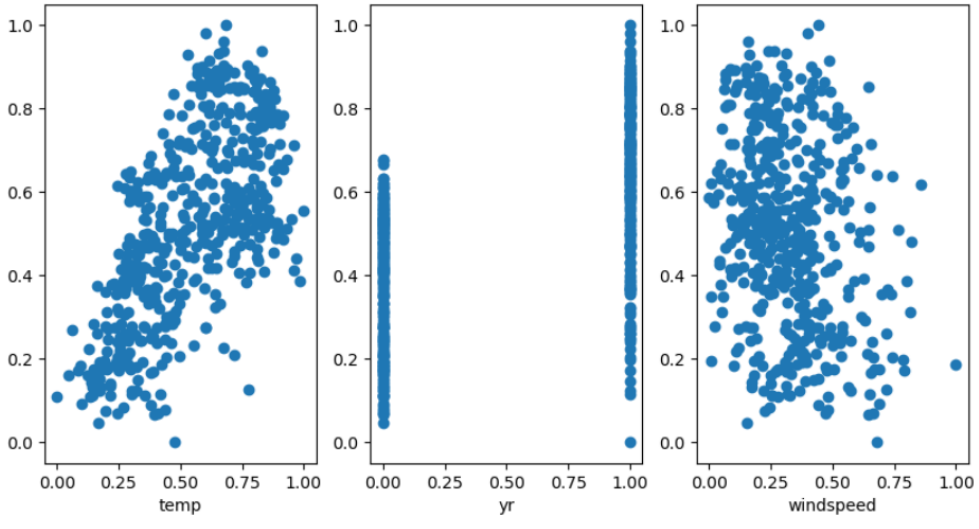
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

A) Ignoring 'registered' & 'casual' which are sub-classification of target variable 'cnt', the highest correlation is with temp & atemp

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

A) A Linear regression model should follow below assumptions:

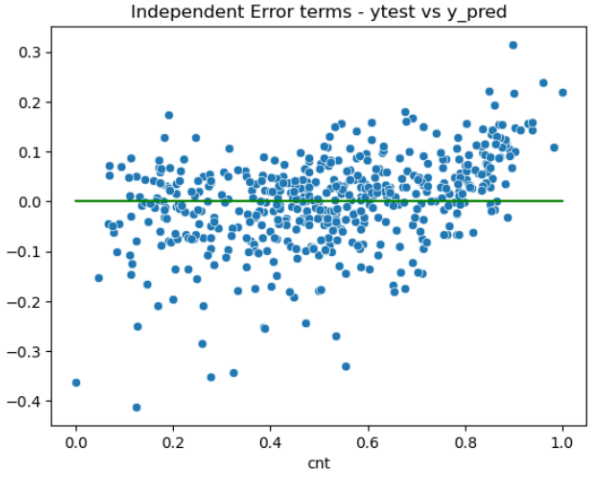
- Linear relation between independent and dependent variables.
    - There should be linearity between independent and dependent variables.



- Error terms are independent:
    - The error terms should be independent to each other and does not follow the time-series like pattern.
    - It can be verified by plotting scatter plot with y_train vs residuals.

```
residual_train = y_train - y_pred
sns.scatterplot(y_train,residual_train)
plt.plot(y_train, (y_train - y_train),'g')
plt.title("Independent Error terms - ytest vs y_pred")
```

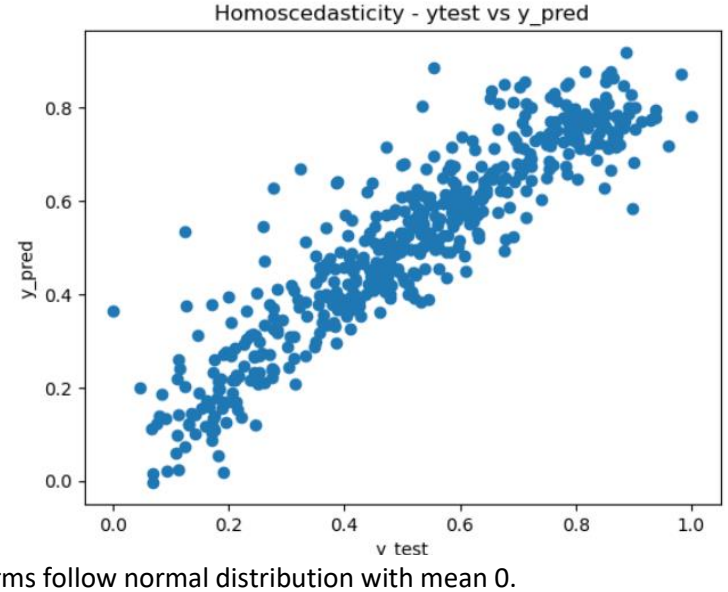
```
Text(0.5, 1.0, 'Independent Error terms - ytest vs y_pred')
```



- Homoscedasticity: Error terms are constant variance
  - The variance should follow and pattern, like sudden increase or decrease.
  - It can be verified by plotting scatter plot btw y_train and y_pred

```
plt.scatter(x=y_train, y = y_pred)
plt.title("Homoscedasticity - ytest vs y_pred")
plt.xlabel("y_test")
plt.ylabel("y_pred")
```

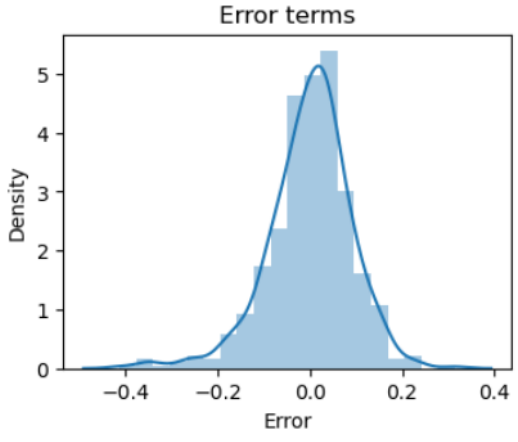
```
Text(0, 0.5, 'y_pred')
```



- Error terms follow normal distribution with mean 0.
  - The error terms which are the residual should have normal distributed with mean 0
  - It can be verified by plotting distplot with residuals(y_train – y_pred)

```
plt.figure(figsize=(4,3))
sns.distplot((y_train - y_pred),bins=20)
plt.title("Error terms")
plt.xlabel("Error")
```

```
Text(0.5, 0, 'Error')
```



- No multicollinearity
  - There should not be high correlation between the dependent variables.
  - It can be verified by checking VIF(variance inflation factor) and also checking the correlation with heatmap.

	col_name	vif_score
2	temp	4.76
1	workingday	4.04
3	windspeed	3.44
0	yr	2.02
9	weekday_6	1.69
4	season_summer	1.57
6	weathersit_2	1.53
5	season_winter	1.40
8	mnth_sep	1.20
7	weathersit_3	1.08

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

A) The variables with highest co-efficient have more contribution towards explaining the demand.

temp	0.549892
yr	0.233139
season_winter	0.130655
mnth_sep	0.097365
season_summer	0.088621
const	0.075009
weekday_6	0.067500
workingday	0.056117
weathersit_2	-0.080022
windspeed	-0.155203
weathersit_3	-0.287090

So, from the model, top 3 feature which contribute are:

- Temperature (+0.549)
- Weathersit_3 (-0.287)
- Year (+0.233)

General Subjective Questions

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

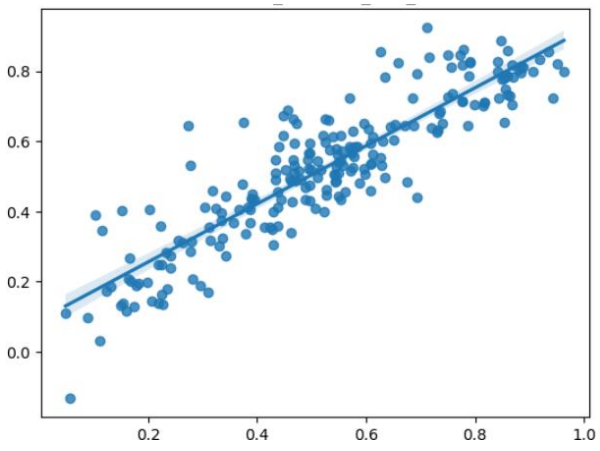
A) Linear regression analysis is a statistical approach in modeling the relation between a target variable (dependent variable) and predictors(independent variable).
Linear regression can be classified:

- Simple linear regression:
  - In this case, there will be only one independent variable and it can be represented as
    $Y = B_0 + B_1 X.,$
    Where ' Y', is target variable and 'X' is independent variable.
    ' B_1' is the co-efficient and ' B_0' is constant

- Multiple linear regression:
  - As name suggests, in case of multiple linear regression, there will be more than one independent variable. Mathematically representation could be

    $$Y = B_0 + B_1X_1 + B_2X_2 + ... + B_pX_p + \varepsilon$$

    Where ' Y', is target variable and 'X' is independent variable.
    ' B_i' is the co-efficient of i^th independent variable
    ' B_0' is contant,
    ε is the error



Note: if the co-efficient is positive, then there is a positive linear relation between target and predictor.

In case of negative, then there is negative linear relation.

Assumptions in linear regression model:

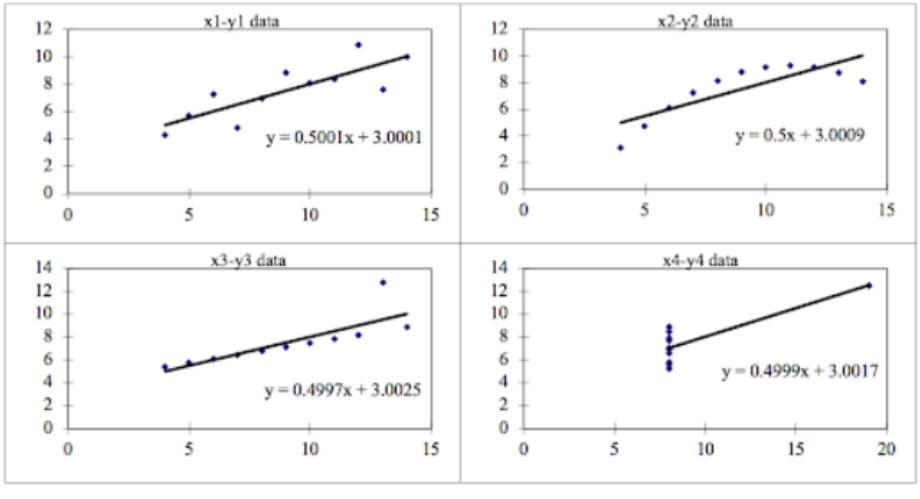
- Linear Relationship:
  - Assumption that there is linear relationship between the target variable and its predictor variables.
- Error terms normality:
  - The error terms when plotted on a distplot, it should be normally distributed with mean at zero.
- Error terms are independent:
  - Error terms should be independent to each other.
  - There should be any kind of visible pattern
- Homoscedasticity:
  - Variance should be contact and there should not be increase or decrease of this value.
  - Also it should not follow any pattern.

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

A) Anscombe's quartet explain about the importance of visualizing the data. In this concept, 4 data sets are taken with identical statistical information.

Anscombe's Data								
Observation	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89
			Summary Statistics					
N	11	11	11	11	11	11	11	11
mean	9.00	7.50	9.00	7.500909	9.00	7.50	9.00	7.50
SD	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94
r	0.82		0.82		0.82		0.82	

From the result of statistic perspective, the usual conception would be that the data distribution is almost similar for all the 4 data sets as the statistical summary is similar.
However, when the data is actually plotted with scatter plot, each dataset generates different kind of plot.



From the plot, we can say:
Dataset -1 fits the linear regression
Dateset -2 shows the non-linear pattern.
Dataset -3 has outliers in the data
Dataset -4 has one outlier with x=19 which disturbs the model

So from the four datasets, we can only use linear regression on Dataset-1.

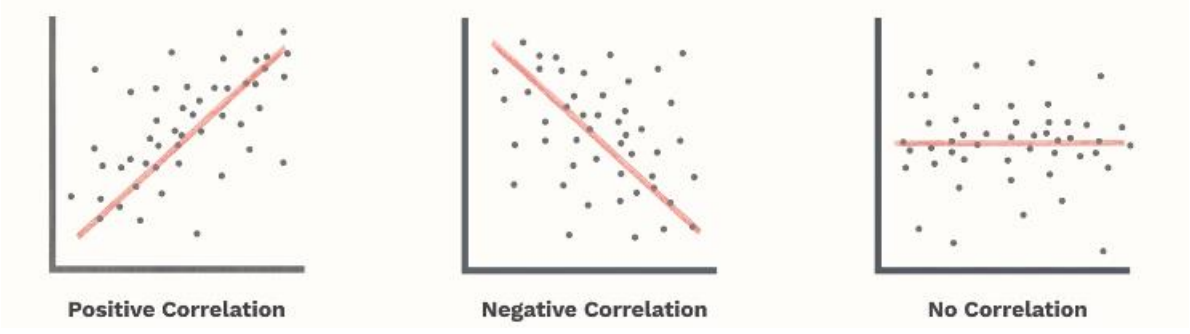
3. What is Pearson's R?
A) Pearson's correlation coefficient(R) is the measure of linear correlation between the variables. Specifically, it describe the strength and direction of linear relationship between two variables.

Its value range from +1 to -1. A value between 0 to 1 indicates a positive relation and the values between -1 to 0 represent negative relation. And value of 0 represents that there is no co-relation exists between the variables.

Positive correlation: As 'X' increase 'y' increase

Negative correlation: As 'X' increase 'y' decrease.

No correlation: No relation exists.



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

A) Scaling is a process which standardize all the independent values in order to bring all the data in a same scale. It helps in handling different units and higher numerical values and convert into lower range. It helps to do calculations in algorithms very quickly and also it takes less time to train the model.

Scaling is done during the data preprocessing. When the feature scaling is not performed, then the machine learning model gives higher weightage to higher value and lower weightage to lower values.

Normalisation	Standardisation
Mathematical expression: $Xn = (X-X_{min}) / (X_{max} – X_{min})$	Xstand = (X – mean(X)) / standard_dev
It uses minimum and maximum values	It uses mean and standard deviation
Range is between 0 to 1 or -1 to 1	It is not restricted to any range
It is used when distribution of the variable is not clear	If the distribution of variable is consistent, then this method is used
After by outlier	Affected is less by outliers
It is also known as scaling normalization	It is also known as Z-score

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

A) The VIF is given by :

$$VIF_i = 1 / (1 – R_i2)$$

VIF is infinite when R2 is equal to 1 which happens when it there is a perfect correlation between two independent variables.

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

A) QQ plots help us identify distribution types by visually comparing data from two different sources onto one plot. This quickly allows us to see if our data follows the tested distribution. A QQ plot can be used to test for a match with any distribution.

Its importance is that the QQ Plot can ensure your data is a correct distribution because, your data and the data from the distribution will match perfectly. If they do not, your data is either from a different distribution, has outliers, or is skewed, altering it off the true theoretical distribution.