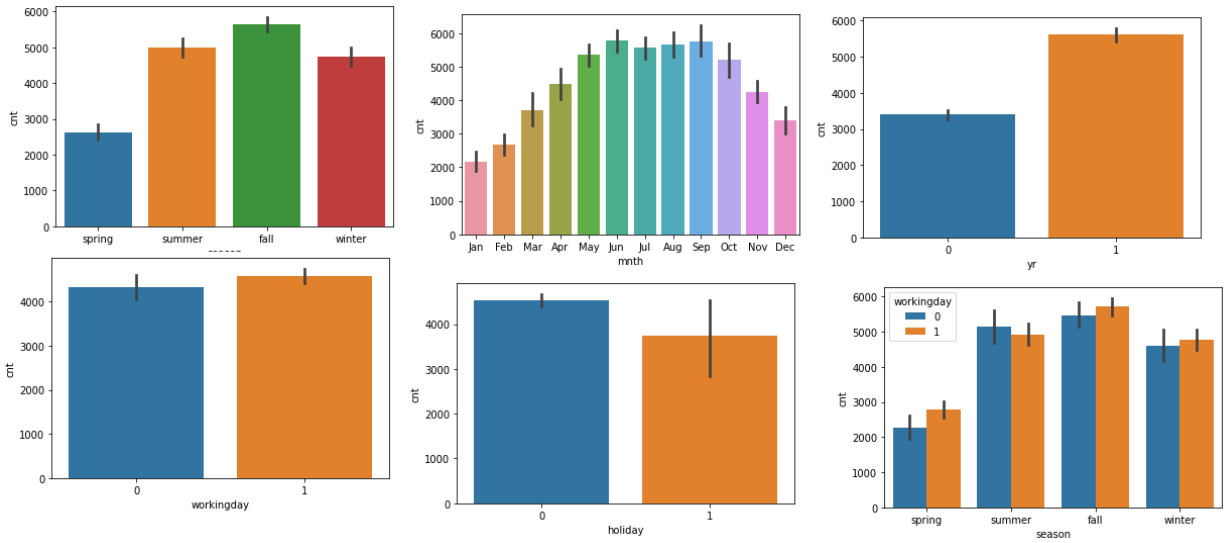
Assignment-based Subjective Questions

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: - Below are some insights we got from categorical variables in the data set .

- ❖ Summer and Fall season have high bike rental count.
- ❖ Better the weather, more the count of rental bikes
- ❖ Bike rental count is increasing from Jan to Jun, remains constant till sept and then decreasing.
- ❖ People rent more bikes in 2019.
- ❖ In all the seasons people has rented bike more in 2019
- ❖ people rent more bike during working day.
- ❖ people rent less bike during holiday.
- ❖ In fall season people rent more bike during holiday and in Summer, Winter and Spring less during holiday.
- ❖ In Summer season people rent more on not working day, while in spring, fall and in winter people rent more bike during working day.



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

Answer:- drop_first parameter is use to create n-1 variable . If we have n categories inside any categorical variable, then n-1 dummy variable should be created to remove the multicollinearity.

A dummy variable is a numerical variable used in regression analysis to represent subgroups of the sample.

e.g. below screenshot

```
#Creating the dummies variable for above categories
brdf = pd.get_dummies(data = bikerentaldf , columns = cat_variables , drop_first=True)
print(brdf.shape)
brdf.head()
```

```
(730, 29)
```

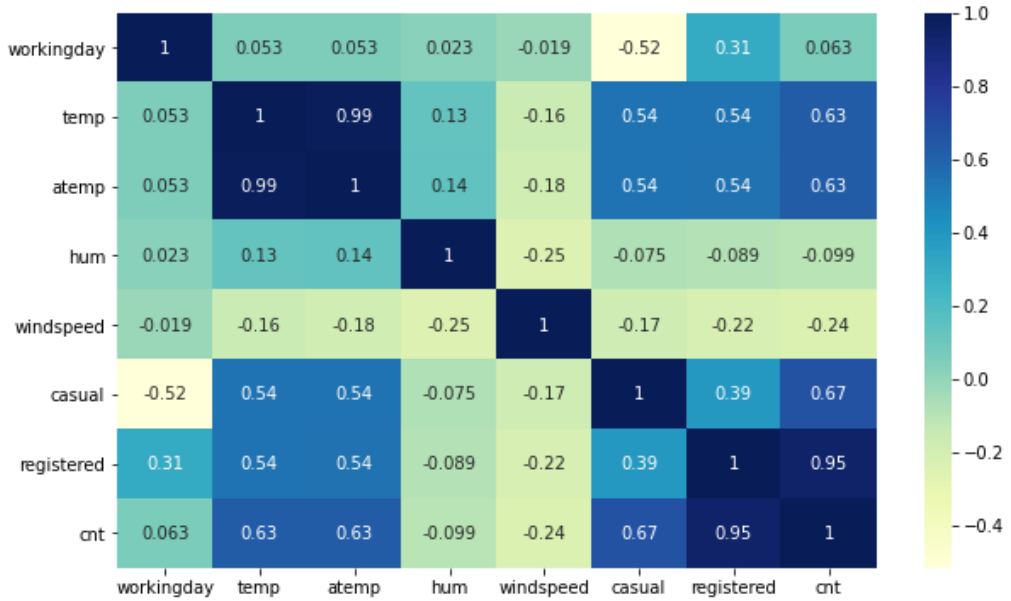
]:

	yr	holiday	workingday	temp	hum	windspeed	cnt	season_summer	season_fall	season_winter	mnth_Aug	mnth_Dec	mnth_Feb	mnth_Ja
0	0	0	0	14.110847	80.5833	10.749882	985	0	0	0	0	0	0	
1	0	0	0	14.902598	69.6087	16.652113	801	0	0	0	0	0	0	
2	0	0	1	8.050924	43.7273	16.636703	1349	0	0	0	0	0	0	
3	0	0	1	8.200000	59.0435	10.739832	1562	0	0	0	0	0	0	
4	0	0	1	9.305237	43.6957	12.522300	1600	0	0	0	0	0	0	

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

Answer: - As we see that count has highest correlation with Registered → casual → temp, atemp.

Registered and casual are also count of users and we have to make our model with total count, hence temp has the highest correlation.



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

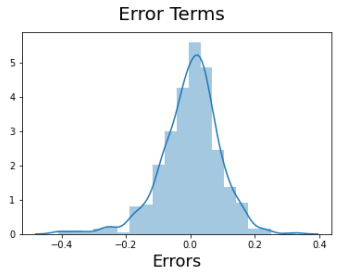
Answer: -

Linear Relationship: It was found that there is a linear relationship found b/w independent and dependent variables.

Multicollinearity: Removed multicollinearity by VIF and p-value method.

Homoscedasticity: It was found that residuals have constant variance.

Normal Distribution of Residuals: Did residual analysis and found normal distribution for error terms.



As we can see error terms are normally distributed

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: - According to final model

- Temp
- Year
- Season_winter

If management team has to decide the top 3 drivers then I think mnth_sep will be considered , they might drop Year because they can increase the production of bikes or provide sufficient bikes to people in September month or in winter season or when the temperature is better.

```
#model coefficients
lr_model_final.params.sort_values(ascending=False)
```

```
temp                        0.552174
yr                          0.233084
season_winter               0.129723
const                       0.128458
mnth_Sep                    0.095875
season_summer               0.089238
weekday_Sunday             -0.046400
weathersit_Mist + Cloudy   -0.078571
windspeed                  -0.155211
weathersit_Light Snow      -0.283276
dtype: float64
```

General Subjective Questions

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

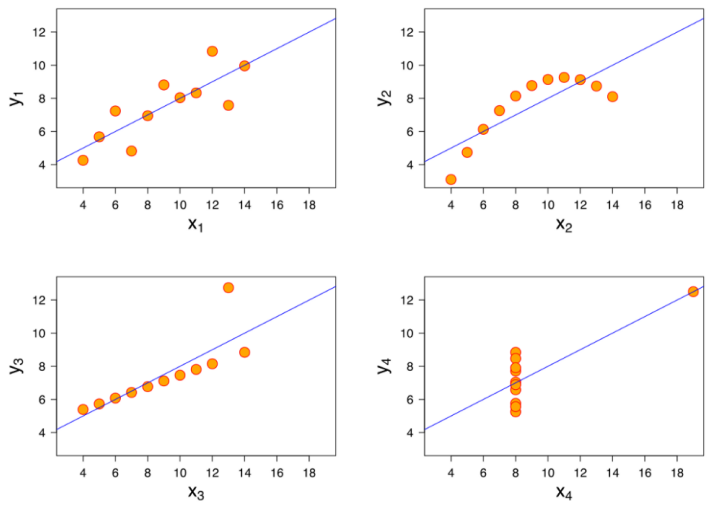
   Answer: The objective of a linear regression model is to find a relationship between one or more features (independent variables) and a continuous target variable (dependent variable). When there is only feature it is called *Uni-variate* Linear Regression and if there are multiple features, it is called *Multi*ple Linear Regression.
   so whenever we find any relationship b/w independent and dependent variable we can apply linear regression.
   Algorithm:
   1. Understand the data set and visualize the patterns, identify target variables and independent variables.
   2. Find the correlation b/w all the continuous variable with target variable.
   3. Perform segmented univariate analysis on all the categorical variables and try to understand their relationship with target variable.
   4. Remove null values if there are any or fill it with mean or median.
   5. Create n-1 dummy variables for categorical columns and convert all the columns to numerical values.
   6. Split the dataset in train and test.
   7. Apply scaling (min-max or standardization) on train set.
   8. Create X_train , y_train from train dataset.
   9. For Simple linear regression we can directly create model here with one independent variable and check the summary
   10. Else for Multiple regression we have 2 options manual backward elimination approach where we have to take all the features and then remove each one by one after checking p-value and VIF or Recursive feature elimination.
   11. If variables are greater than 10 then we can use RFE , where we have to pass the number of top predictors we have to find.
   12. Once we got top predictors from step 11 , Create model and check its summary
   13. Check the p-value and VIF and remove the features having high p-value (>0.05) also high VIF (>5) .Make sure to use below priority while removing features.
       - High p-value and high VIF
       - High p-value and low VIF
       - Low p value and high VIF
   14. Continue step 13 until we got all significant features (p-value<0.05) and VIF<5. Make sure to have good R-Square and adjusted r -squared and F statistic (close to zero means overall model is fit)
   15. Make predictions which is y_train_pred values and perform residual analysis on final model.
   16. Plot the residuals to check normal distribution and constant variance.

17. Scale test set, make sure only to transform and not fit
18. Make prediction on test set.
19. For model comparison check the Adjusted r squared of test set.
20. If there is less than 5 % difference b/w R-Squared and Adjusted r squared of test and train set then it means it's a good model
21. Sort the drivers by coefficients to provide the result to management team to take business strategic decisions.

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

Answer: - Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x,y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely when we plot these datasets. Each graph tells a different story irrespective of their similar summary statistics.

- Let's say you have 4 data set and he summary statistics show that the means and the variances and standard deviation were identical for x and y across the groups.
- The correlation coefficient (how strong a relationship is between two variables) between x and y is also same for each dataset.
- When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well, but each dataset is telling a different story:



- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

Hence This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

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

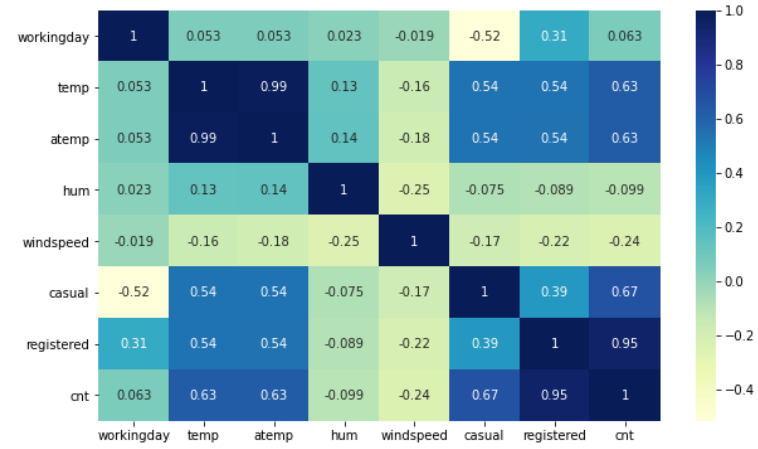
Answer:-

This is used to find out the correlation between independent and dependent variables.

Pearson's r is a numerical summary of the strength of the linear association between the variables. Let's say x and y are 2 variables if increase of x, y is also increasing then correlation coefficient will be positive. if increase of x, y is decreasing, the correlation coefficient will be negative.

In Python we have "corr()" function to calculate the correlation coefficient.

```
plt.figure(figsize=[10,6])
sns.heatmap(bikerentaldf[ ['workingday', 'temp' ,'atemp' ,'hum' ,'windspeed' ,'casual' ,'registered','cnt']].corr() ,annot=Tr
            ,cmap = 'YlGnBu')
plt.show()
```



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

Answer: Scaling is required to transform the data set in specific scale either 0-1 or 0-100. scaling doesn't impact your model. It is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation. So it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale. As you know, there are two common ways of rescaling:

The most popular methods for scaling:

- Min-Max Scaling – Normalization
  - It converts all the values in 0 and 1.
  - x-min(x)/max(x)- min(x)
- Standard Scaling – standardization

- With this all the data will be having mean of 0 and standard deviation of 1
- X-mean(x)/sd(x)

The difference is that, in scaling, you're changing the range of your data while in normalization you're changing the shape of the distribution of your data.

why scale?

- Helps with interpretation.
- Faster convergence of gradient descent

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

   Answer: - A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

   1. This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation).
   2. The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity.
   3. A general rule of thumb is that if VIF > 10 then there is multicollinearity. Note that this is a rough rule of thumb, in some cases we might choose to live with high VIF values if it does not affect our model results such as when we are fitting a quadratic or cubic model or depending on the sample size a large value of VIF may not necessarily indicate a poor model.

   It is calculated by:-

   VIF_=1/(1-R-Squared^2)

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
   Answer: - : The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.