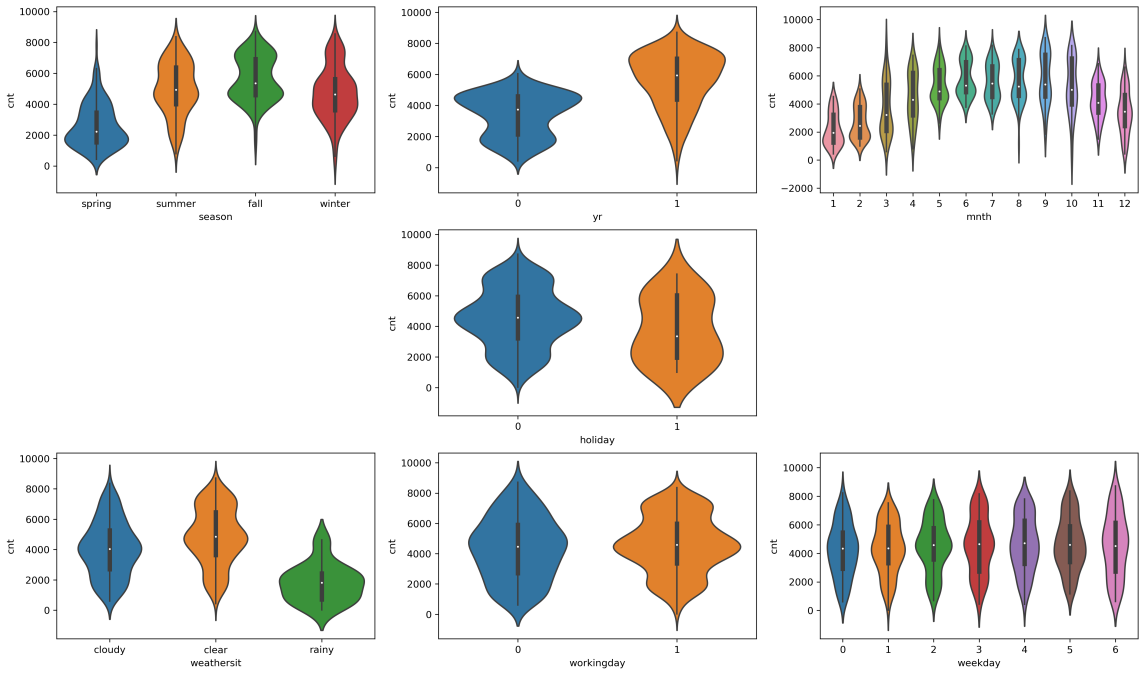
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? **[3 mark(s)]**

    - Below is the violin plot showing the distribution of demand across several categorical variables.



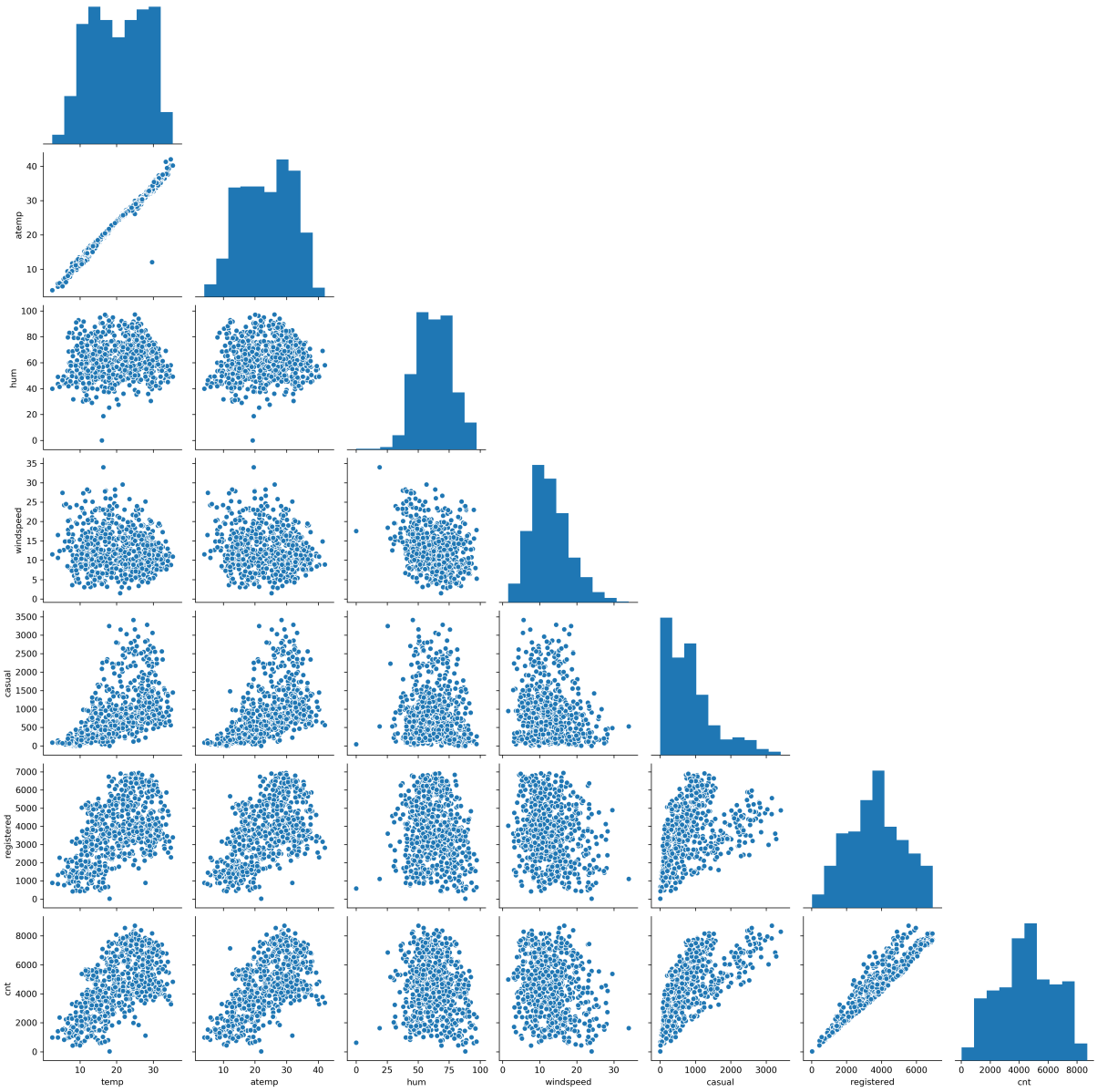
    - **Insights:** (**MCBR** - Median count of Bikes Rented)

        - Spring and Winter season has the lowest MCBR amongst all the seasons. Usually the road conditions in spring are worse due to the previous winter and will start getting repaired. However, we do have some adventurous outliers that are thrill seekers and understand the risks.
        - MCBR has increased much more in 2019.
        - MCBR steadily increases from January till July. Then starts to drop from July to December. This goes along with the same insight that we got from the seasons.
        - MCBR is higher for clear and cloudy compared to rainy conditions. Clear weather being the highest for bike demands.

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

- E.g. here we have seasons as **spring, summer, fall and winter**.
- The dummy variable encodings corresponding to them would have values `0001, 0010, 0100, 1000`.
- This would add 4 variables to the set of independent variables. Which would increase the complexity of the model.
- This same information can be endcoded as `001, 010, 100` with `000` signifying **spring** season.
- This would help us capture the same information without adding to the complexity of the model.

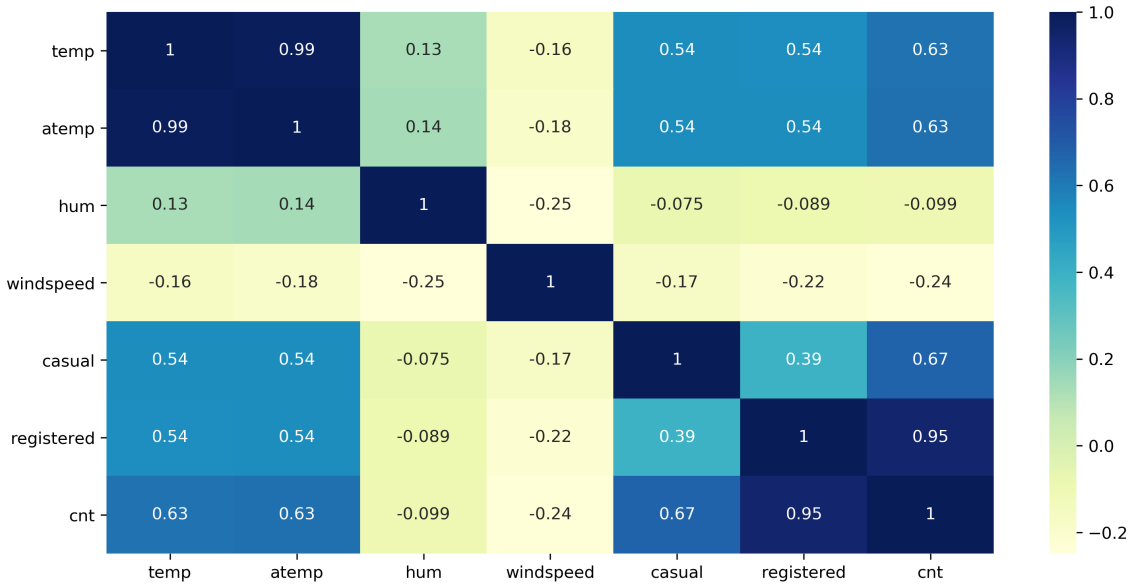
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? **[1 mark(s)]**



- Looking at these pair-plots. We can see that `registered` and `casual` are

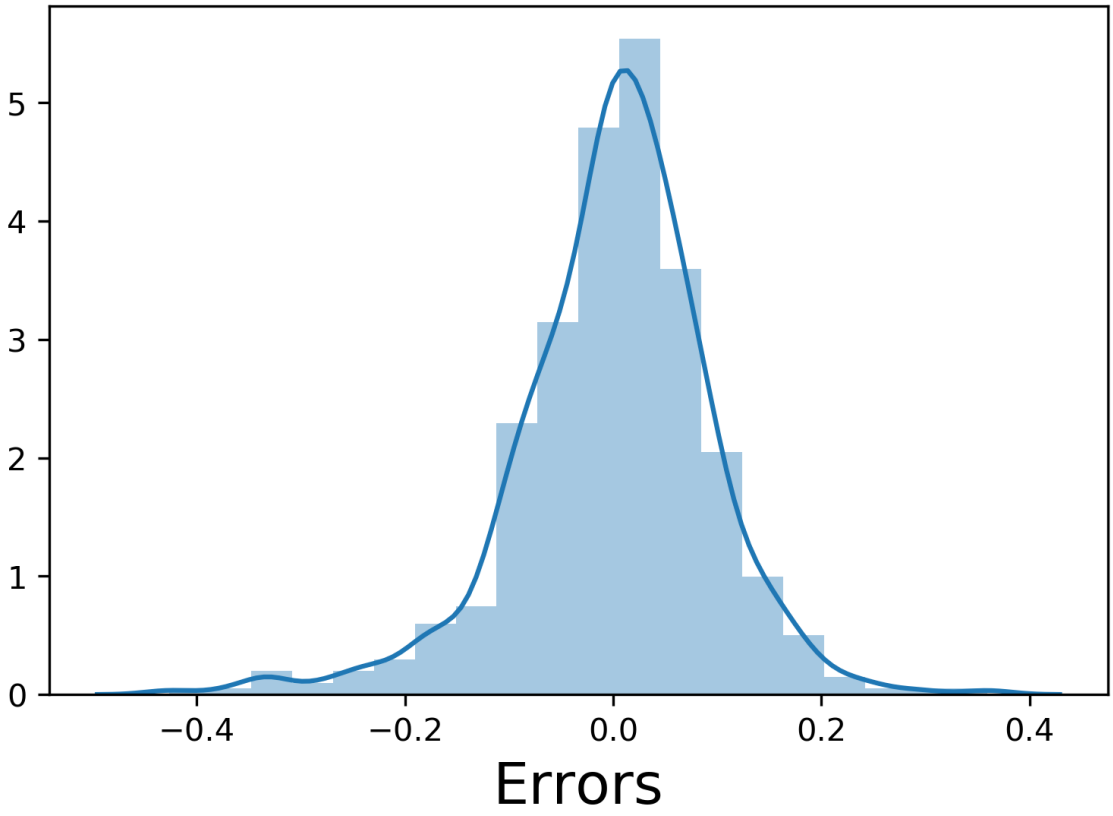
highly correlated with `cnt`. But, this is because they all represent the same thing, the demand of bikes. So, we won't include them in building our model. Because these are our target variables.

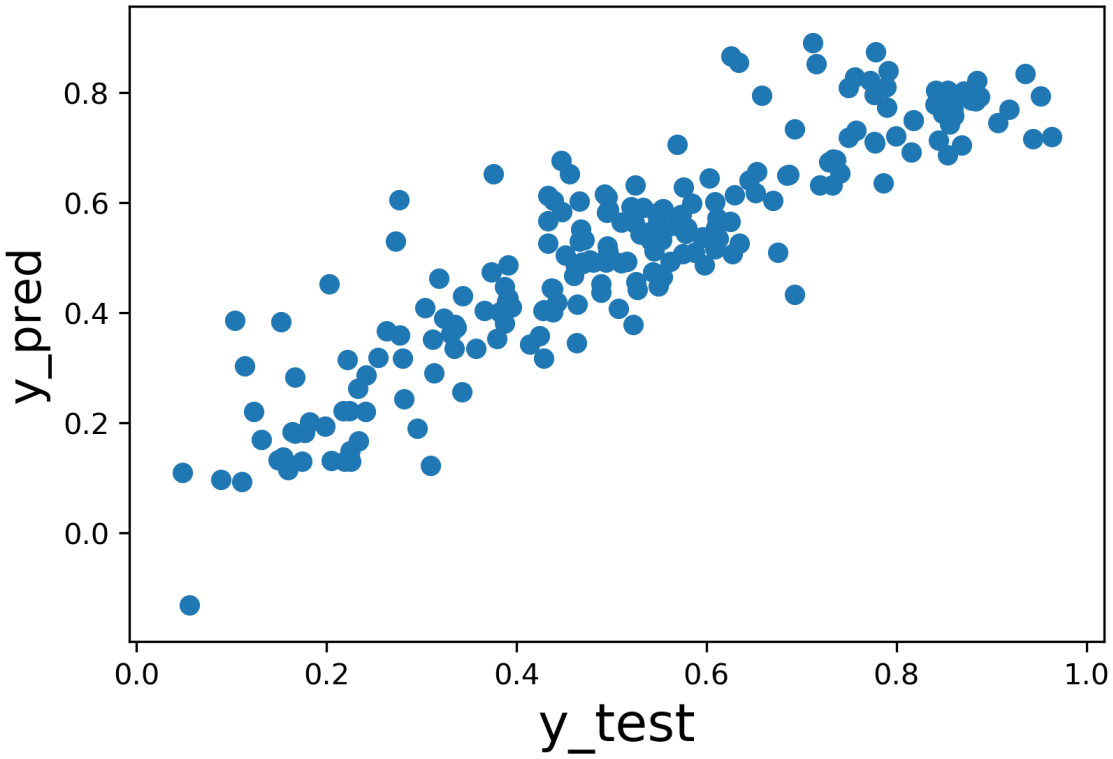
- So, from the plots `temp` has the highest correlation with our target variable.
- Below, is the correlation heatmap for the same.



4. How did you validate the assumptions of Linear Regression after building the model on the training set? **[3 mark(s)]**

    - We validated our assumptions using **Residual Analysis** on the training data and visualizing the distribution of error terms. We found that the distribution was normal which validated our training model.
    - Further, we used the model to predict the demand on the test data set and found the correlation between `y_test` and `y_pred`.
    - We also calculated the `r2_score` between `y_test` and `y_pred`. It turned out to be `0.8009724887482659`.





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

- Top 3 features contributing towards explaining the demand are temperature, holidays, rainy weather and windspeed.
- Below is the summary of the the trained model.

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.823
Model:	OLS	Adj. R-squared:	0.820
Method:	Least Squares	F-statistic:	258.0
Date:	Sun, 30 Aug 2020	Prob (F-statistic):	1.42e-181
Time:	12:56:45	Log-Likelihood:	479.80
No. Observations:	510	AIC:	-939.6
Df Residuals:	500	BIC:	-897.3
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2223	0.030	7.436	0.000	0.164	0.281
yr	0.2340	0.009	27.430	0.000	0.217	0.251
holiday	-0.0877	0.027	-3.254	0.001	-0.141	-0.035
temp	0.4667	0.034	13.720	0.000	0.400	0.533
windspeed	-0.1546	0.026	-5.948	0.000	-0.206	-0.104
spring	-0.0824	0.021	-3.947	0.000	-0.123	-0.041
summer	0.0371	0.014	2.648	0.008	0.010	0.065
winter	0.0760	0.017	4.512	0.000	0.043	0.109
cloudy	-0.0763	0.009	-8.438	0.000	-0.094	-0.059
rainy	-0.2794	0.026	-10.919	0.000	-0.330	-0.229

Omnibus:	62.627	Durbin-Watson:	2.015
Prob(Omnibus):	0.000	Jarque-Bera (JB):	154.655
Skew:	-0.635	Prob(JB):	2.61e-34
Kurtosis:	5.380	Cond. No.	16.9

General Subjective Questions

1. Explain the linear regression algorithm in detail. **[4 mark(s)]**

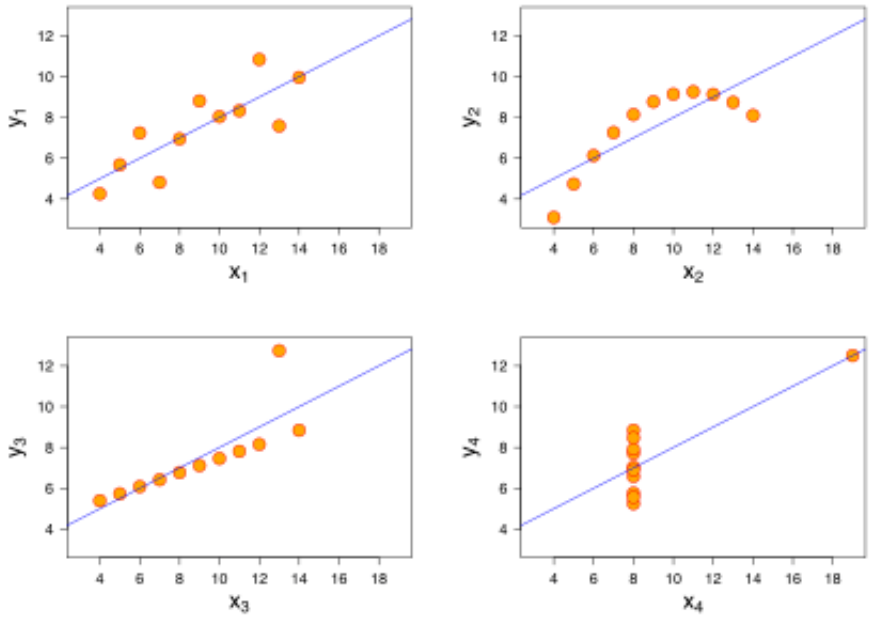
   - Linear regression is a Supervised learning algorithm.
   - We can use it to predict targets given we have past data. For example:
     - predicting the house prices given we have housing dataset
     - predicting the demand of bikes given the past data of bike rentals

   - Assumptions of linear regression are:
     - It assumes linear relationship between target variable and the independent variables.
     - Error terms are normally distributed with mean 0.
     - Error terms are independent of each other.
     - Error terms have constant variance.

- Making these assumptions, the goal of Linear regression is to find the best-fit line that can explain the variance in the data.
- It does so by minimizing the residual sum of squares (RSS).

2. Explain the Anscombe's quartet in detail. **[3 mark(s)]**

   - Anscombe's quartet is a collection of 4 data-sets.
   - It was created to magnify the importance of data-visualization for analysis and effect of outliers on statistical properties as all 4 data-sets have same descriptive statistics (mean, median, mode, etc.).
   - Below is an image showing all 4 datasets. (source - Wikipedia)



3. What is Pearson's R? **[3 mark(s)]**

   - Pearson's R or Pearson Correlation Coefficient measures linear correlation between two variables.
   - This correlation defines the dependence of both the variables on each other.
   - A high positive correlation means that if one increases the other will also increase and vice versa.
   - A high negative correlation means that if one increases the other will decrease and vice versa.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? **[3 mark(s)]**

   - Scaling of features defines the process of bringing down all the independent variables

in a comparable scale.
    - If the independent variables are in different scales, their coefficients would also be in different scales. And, it would be very hard to find out which independet variable has a higher impact on predicting the target variable.
    - This scaling is performed so that we can interpret and compare the relative significance of the independent variables in predicting the target variable. Also, the the Gradient Descent Algorithm takes lesser time to converge if the independent variables are on the same scale.
    - Normalized scaling brings every variable between a fixed range **([0, 1], [-1, 1])** whereas, standardized scaling distributes them across their mean (standard deviation would become 1).
    - Normalization will handle outliers and compress the data between a range. Standardization doesnt do any such thing.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? **[3 mark(s)]**

    - An infinite VIF value signifies that the independent variable can be represented exaclty by a linear combination of other independent variables.
    - This can happen if we have variables that are highly correlated to each other in our independent variables list.

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