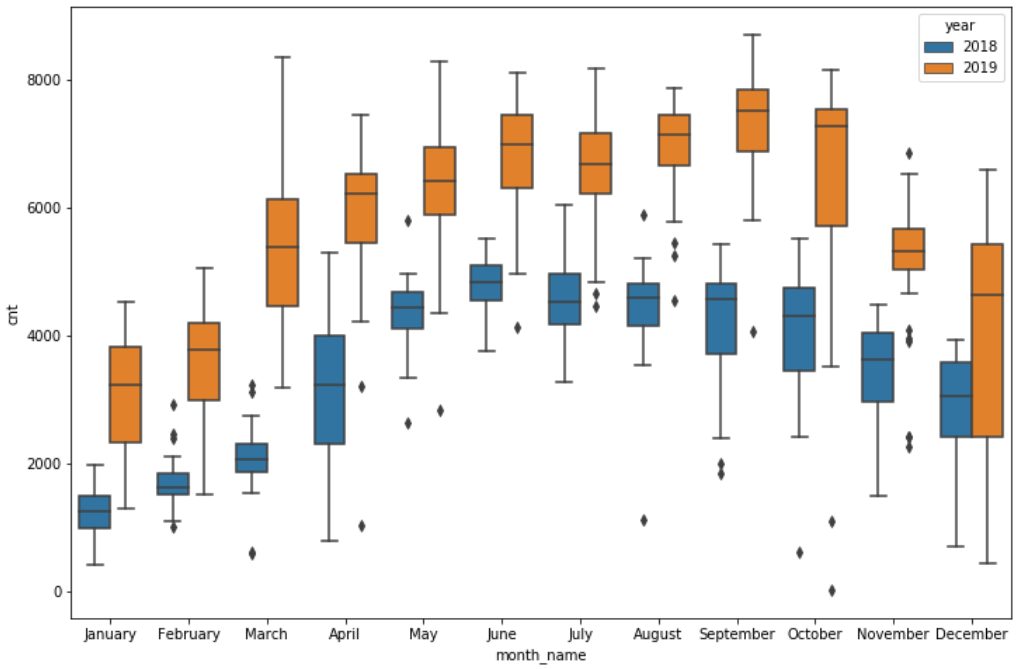
Assignment-based Subjective Questions (by - Sridhar.Chetan@gmail.com)

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

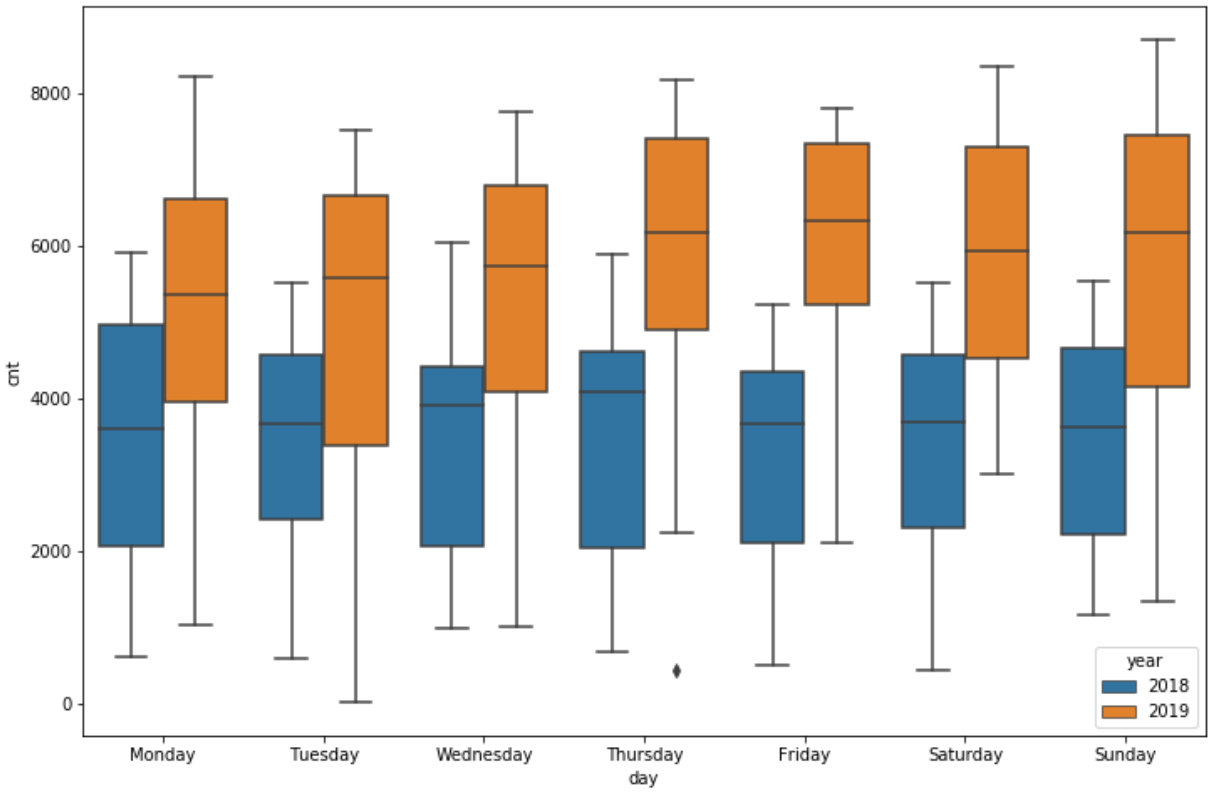
- **Categorical variables: Months, Season, Weather Condition , Day Of Week etc**
- **Target Variable = cnt (rider count)**

➢ **Months vs cnt:**

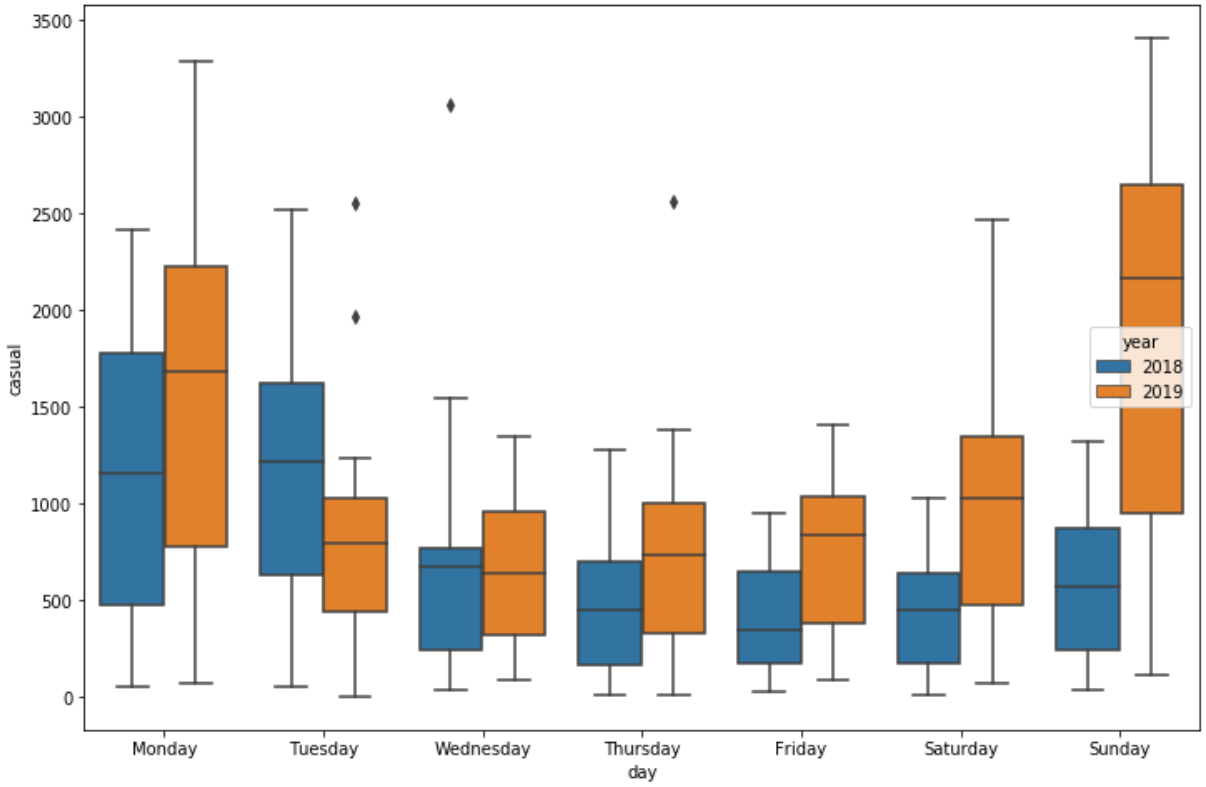
- *Across both years the count increased from Jan to December. The increase is significant during the second and third quarters.*
- *User base seems to have been risen in 2019 compared with 2018.*

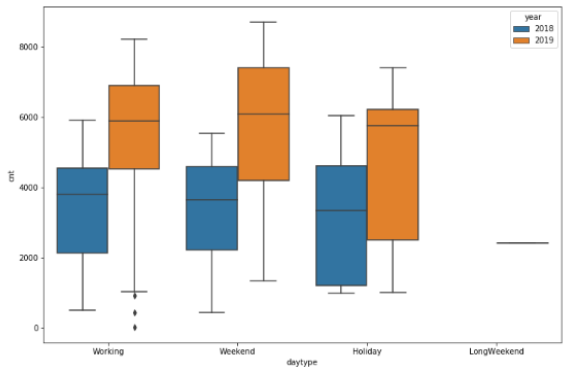


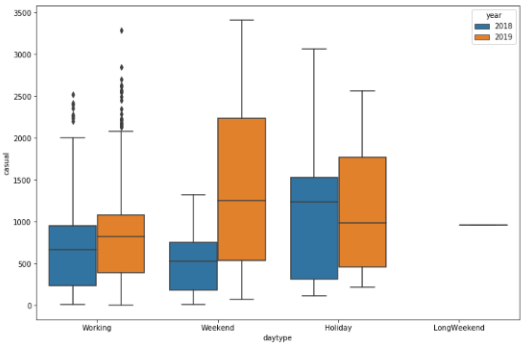
➢ **Day of week vs cnt :**



- *The count is spread across week days with slight increase observed on Thursday and Friday.*
- *Casual users tend to show higher ridership on Monday and Sundays. Significant increase in the median across all other days.*



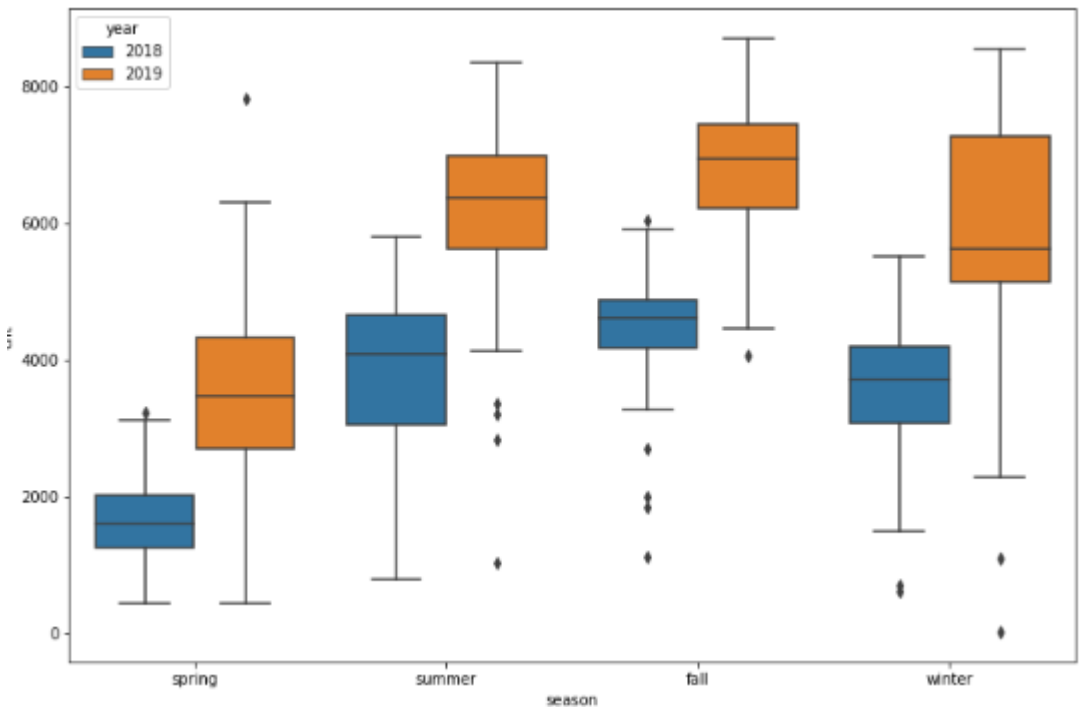




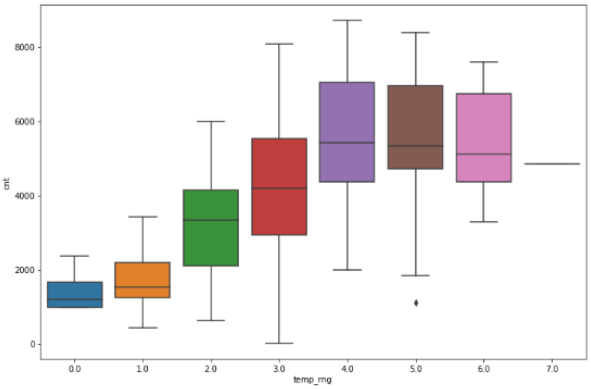
- *Casual users tend to show higher ridership on weekends more compared to registered users.*

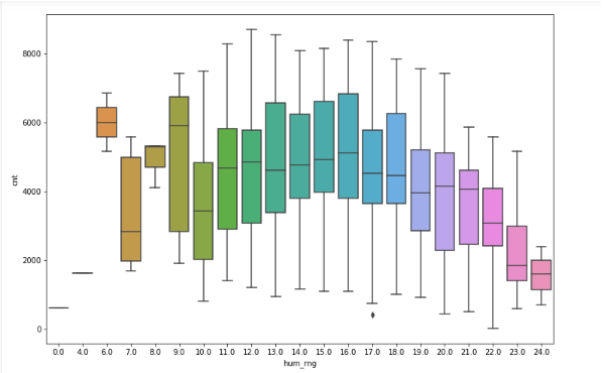
➢ **Seasons vs cnt :**

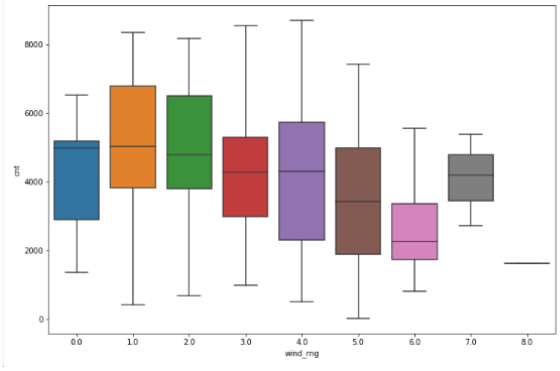
- *Spring shows the least riders count whereas summer and fall show maximum ridership.*
- *There is increase observed in 2019 winter as compared to winter of 2018*



Count vs Temp Range, Wind Range and Humidity







- *Lower Temp ranges show lower ridership. Optimal temp ranges to get the riders attached seems to be around 20-30 degrees.*

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

- *For categorical variables (N Labels) when converted to dummies we get N Columns with each indicating true and false (0/1) as per the category. We need all categorical values fed as 0/1 in ML models as it doesn't work work with Cat data as such.*
- *Consider categorical variable seasons with following values.*

```
In [4]: df.season.value_counts()

Out[4]: fall      188
        summer    184
        spring    180
        winter    178
        Name: season, dtype: int64
```

```
In [12]: season_= pd.get_dummies(df.season,drop_first=True)

In [13]: season_.head()
```

Out[13]:

	spring	summer	winter
0	1	0	0
1	1	0	0
2	1	0	0
3	1	0	0
4	1	0	0

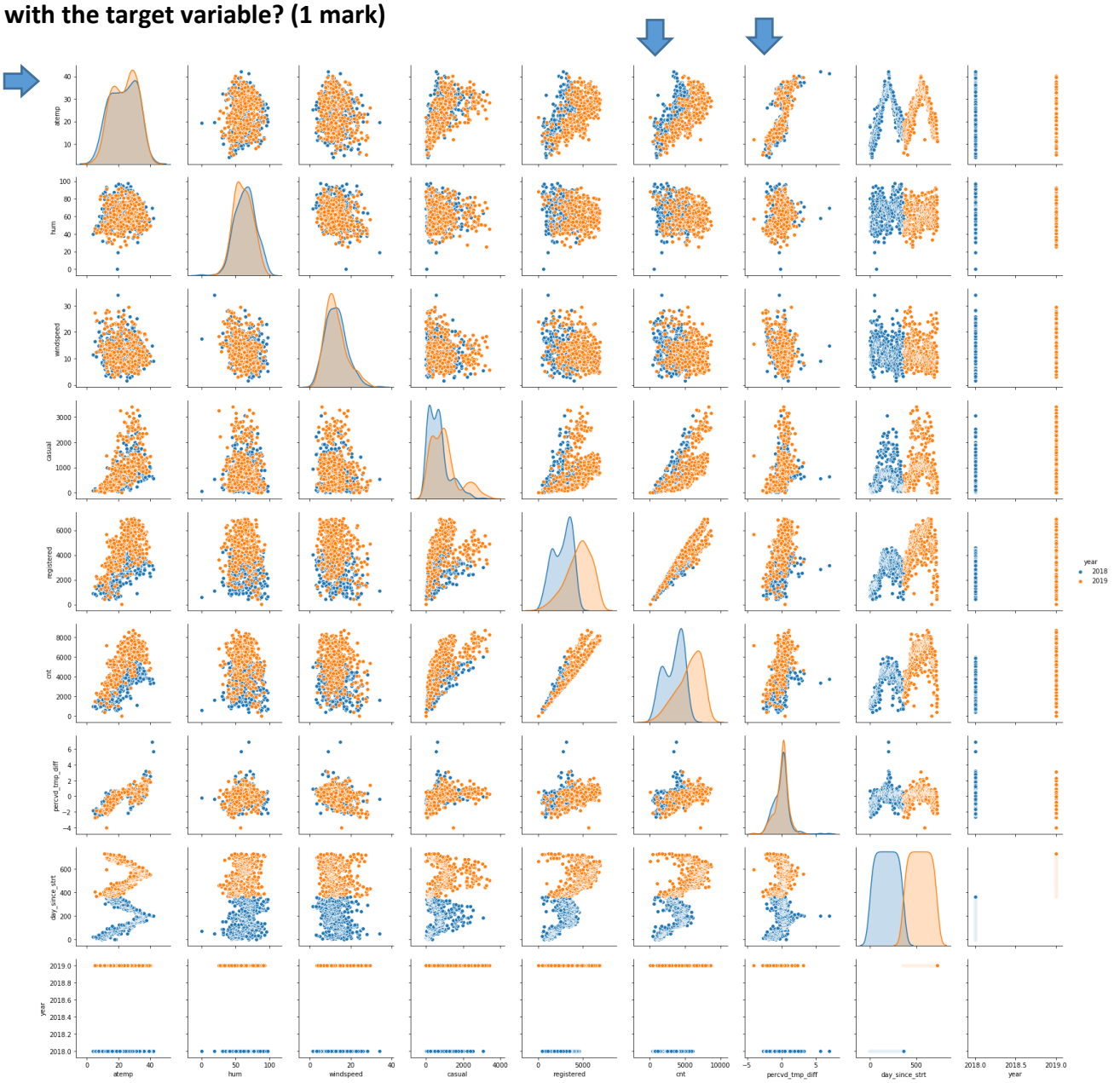
```
In [9]: season_= pd.get_dummies(df.season)

In [10]: season_.head()
```

Out[10]:

	fall	spring	summer	winter
0	0	1	0	0
1	0	1	0	0
2	0	1	0	0
3	0	1	0	0
4	0	1	0	0

- *For analysis without losing any data for modelling we just need N-1 columns hence drop_first = True helps drop the first of the categorical column created (which is redundant). This also helps reduce the number of variables to be fed to the ML model.*

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

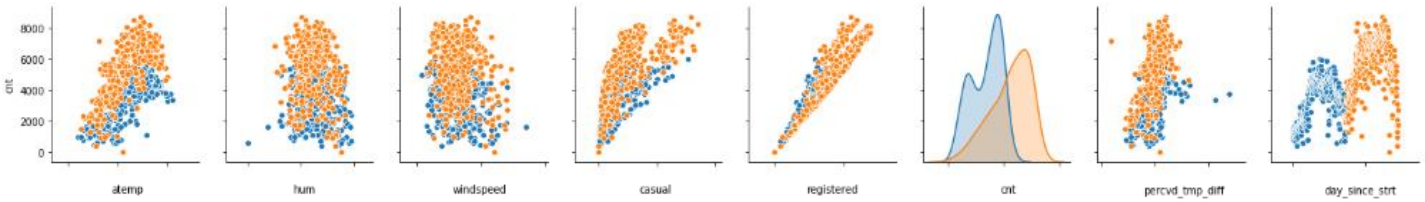


- *Count is strongly related to aTemp*
- *Correlation coefficient is + 0.65*

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

1. The Dependent variable and Independent variable must have a linear relationship.

Pair Plots across numerical variables to find the relationship. Linear relationship was clearly visible across:



- *Linear relationship between Temp , atemp , Percvd_temp_diff ,day_since_start*

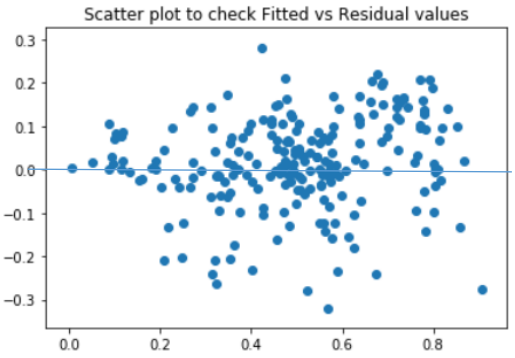
2. No Heteroscedasticity.

Residual vs Fitted values plot can tell if Heteroskedasticity is present or not.

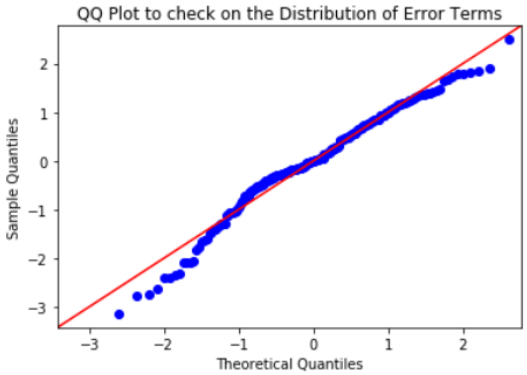
Fitted vs. residuals plot to check homoscedasticity

Variance of the residuals increases with response variable magnitude.

```
93]: x=y_pred_
     y= y_test_ - y_pred_
     plt.scatter(x,y)
     plt.title('Scatter plot to check Fitted vs Residual values')
     plt.show()
```



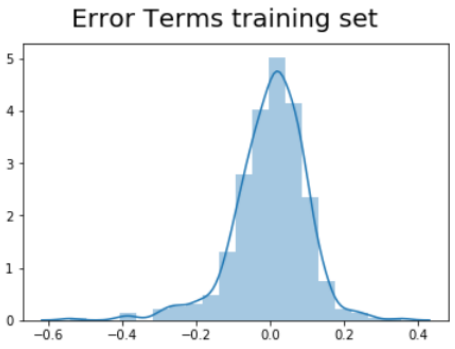

```
In [2818]: Serr = y_test_ - y_pred_
           sm.qqplot(Serr,fit=True, line='45')
           plt.title('QQ Plot to check on the Distribution of Error Terms')
           plt.show()
```



3. Residuals must be normally distributed.

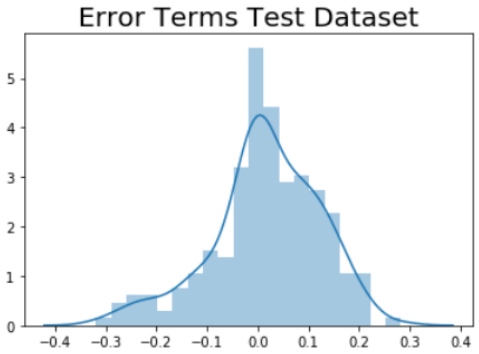
In [2717]:
```
# Residual Error Analysis
y_train_pred_ = lm2_.predict(X_train_lm02)
fig = plt.figure()
sns.distplot((y_train_ - y_train_pred_), bins = 20)
fig.suptitle('Error Terms training set ', fontsize = 20)
```

Out[2717]: Text(0.5, 0.98, 'Error Terms training set ')




```
# Making predictions
y_pred_ = lm2_.predict(X_test_new)
```

```
plt.figure(figsize=(6,4))
sns.distplot((y_test_ - y_pred_), bins = 20)
plt.title('Error Terms Test Dataset', fontsize = 2
plt.show()
```



4. No Perfect Multicollinearity

Out[2776]:

	Features	VIF
0	day_of_yr	12.81
2	atemp	9.16
1	day_since_strt	5.44
5	winter	3.92
3	windspeed	3.91
6	Clear	2.72
4	spring	1.71
7	Weekend	1.36
8	July	1.33
9	September	1.26

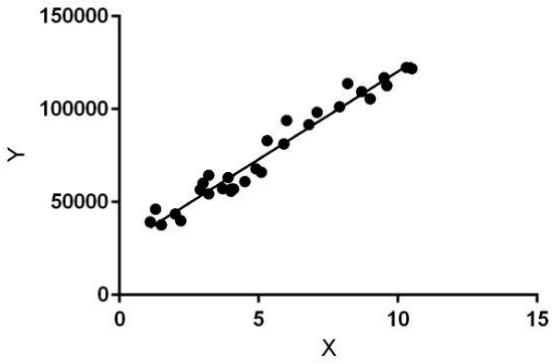
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

	coef	std err	t	P>|t|	[0.025	0.975]
day_since_strt	0.4748	0.018	26.842	0.000	0.440	0.510
atemp	0.4810	0.035	13.643	0.000	0.412	0.550
windspeed	-0.1498	0.027	-5.563	0.000	-0.203	-0.097

General Subjective Questions

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

Linear Regression *is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.*



- *Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression. In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.*
- ***Hypothesis function for Linear Regression :***
  $y = \theta_1 + \theta_2.x$
- *While training the model we are given :*
  x: *input training data (univariate – one input variable(parameter))*
  y: *labels to data (supervised learning)*
- *When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ_1 and θ_2 values.*
  θ_1: *intercept*
  θ_2: *coefficient of x*
- *Once we find the best θ1 and θ2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.*
- **Cost Function (J):**
  By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the θ1 and θ2 values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).

$$minimize \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

Cost function (J) of Linear Regression is the ***Root Mean Squared Error (RMSE)*** *between predicted y value (pred) and true y value (y).*

Gradient Descent:
To update θ1 and θ2 values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line the model uses Gradient Descent Algorithm.

The idea is to start with random θ1 and θ2 values and then iteratively updating the values, reaching minimum cost.

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

Anscombe's quartet *comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.*

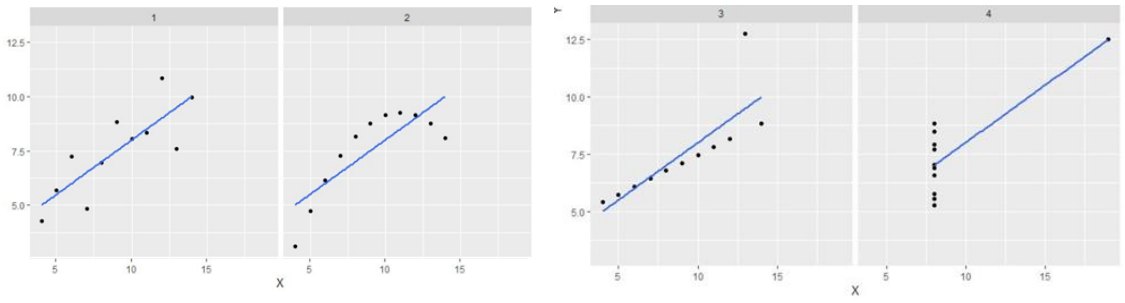
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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

Statistical Summary

Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

For regression line



Explanation of this output:

- *In the first one 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 if you look at this figure you can conclude that there is a non-linear relationship between x and y.*
- *In the third one 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 shows an example when one high-leverage point is enough to produce a high correlation coefficient.*

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

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

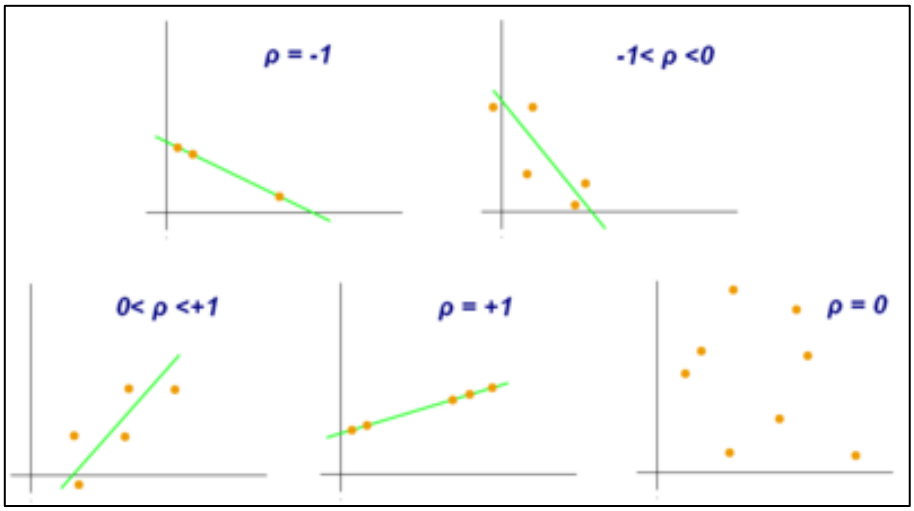
Pearson's r

Correlation is a statistical measure (expressed as a number) that describes the size and direction of a relationship between two or more variables. A correlation between variables, however, does not automatically mean that the change in one variable is the cause of the change in the values of the other variable.

When one variable goes up, does the other go down? Pearson's r can help us with that. Pearson's r is a statistic that helps understand the strength of the linear relationship between two variables.

When r is…

- *Close to 1, there is a strong relationship between your two variables.*
- *Close to 0, there is a weak relationship between your two variables.*
- *Positive (+), as one variable increases in value, the second variable also increases in value. This is called a positive correlation.*
- *Negative (-), as one variable increases in value, the second variable decreases in value. This is called a negative correlation.*



Given a pair of random variables (X,Y), the formula for ρ:

$$\rho_{xy} = \frac{Con\,(r_x, r_y)}{\sigma_x \sigma_y}$$

ρ_{xy} *Correlation between two variables*

$Con(r_x, r_y)$ *Covariance of return X and Covariance of return of Y*

σ_x *Standard deviation of X*

σ_y *Standard deviation of Y*

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

Scaling is a mechanism to transform the numerical variables such that all the dataset features are within a similar range. If the dataset consists of attributes with similar measure and ranges to each other then there's no real need to standardize/normalize.

If, however, some features naturally take on values that are much larger/smaller than others then we do scaling which are of two types: 1) normalization 2) standardization

1. ***Normalization*** *transforms your data into a range between 0 and 1*

   Normalization = x -xmin/ xmax – xmin

2. ***Standardization*** *transforms your data such that the resulting distribution has a mean of 0 and a standard deviation of 1*

   Standardization = x - xmean/ xstd

Normalization/standardization *are designed to achieve a similar goal, which is to create features that have similar ranges to each other. We want that so we can be sure we are capturing the true information in a feature, and that we dont over weigh a particular feature just because its values are much larger than other features.*

Ex If dataset consists of various numerical measures viz - mm , km , Kg , Degrees etc and we pass this to a ML model running algo say Gradient Dissent then the coefficients will be difficult to interpret due to various ranges in the data.

Also – in this case given the ranges and variation in the attributes measures the no of iterations required by the algo to converge to a result will also be considerably higher.

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

Variance Inflation Factor (VIF) measures the severity of multicollinearity in regression analysis. It is a statistical concept that indicates the increase in the variance of a regression coefficient as a result of collinearity.

$$VIF_i = \frac{1}{1 - R_i^2}$$

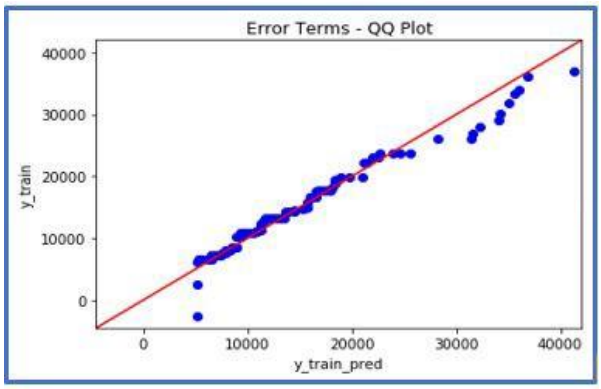
For VIF for a predictor X(i) against the set of predictors is said to infinite when the Ri(2) obtained is = 1 .this is when the variance of X(i) is 100% explained. This would happen that the variable is perfectly correlated.

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or uniform

distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.



A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
Below are the possible interpretations for two data sets.

- Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
- Y-values < X-values: If y-quantiles are lower than the x-quantiles.
- X-values < Y-values: If x-quantiles are lower than the y-quantiles.
- Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x –axis

---