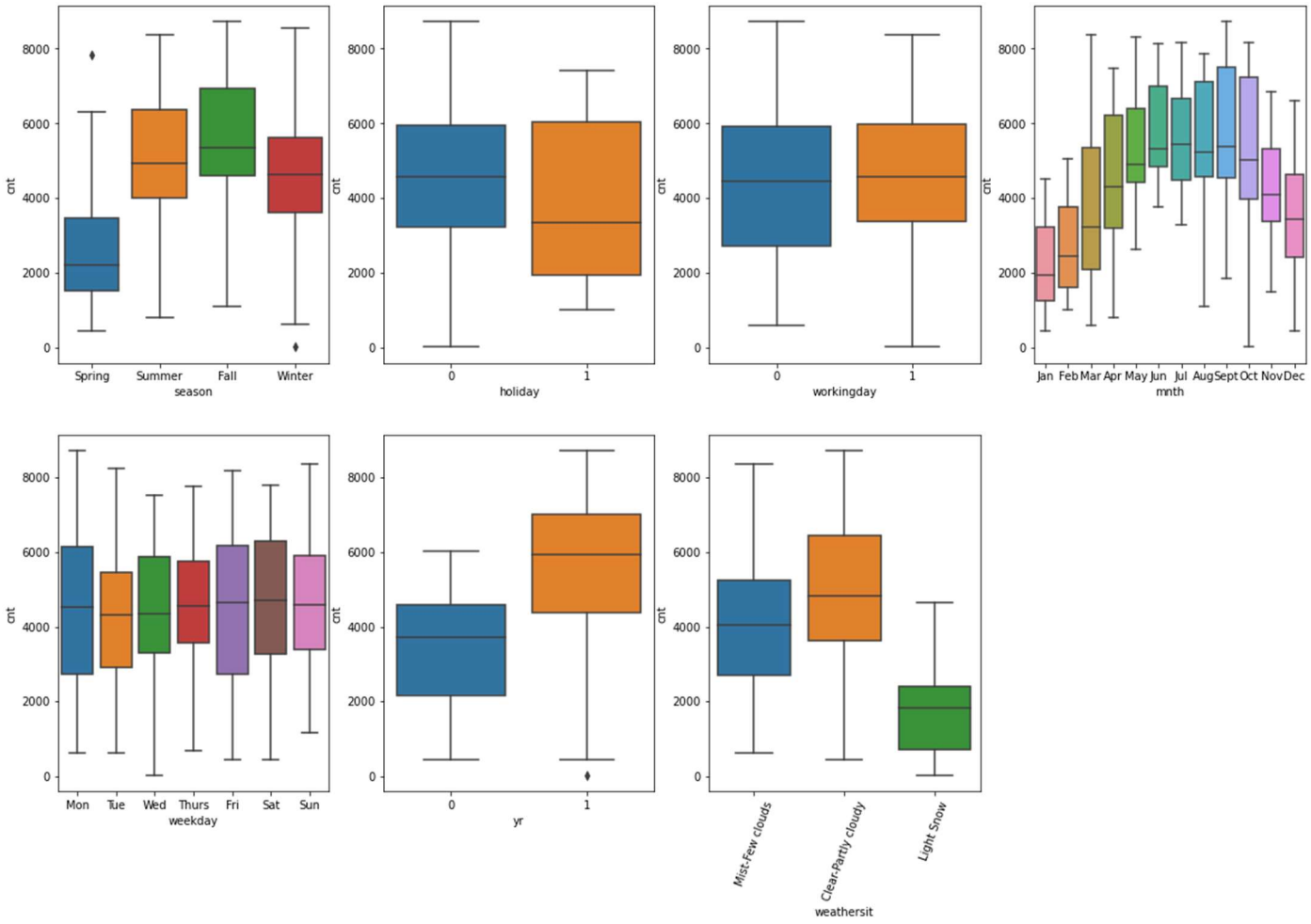
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

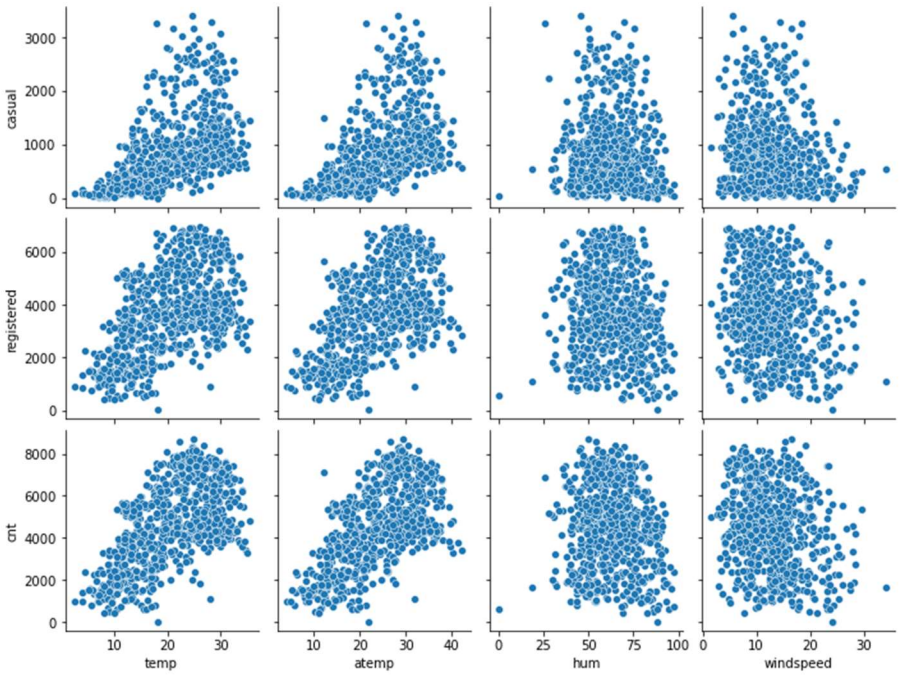
→Bike rental count plotted against Categorical variables as shown below. Some categorical variables have clear corelation with “count” for some can not be inferred from this plot.

1. Seasons "Summer" & "Fall" have somewhat more bike borrowers
2. On holidays bike riders reduce
3. Working day has no much impact on total count
4. May to Sept has comparatively higher demands in year
5. Year 2019 seen higher number of bikers than 2018
6. Clear weather has highest bike rentals while Cloudy have less rentals. Light Snow condition has further deep in rentals.



2. Why is it important to use drop_first=True during dummy variable creation?
   If there are “n” categorical variables, n-1 dummy variables would be created. As 1 variable condition would be redundant and can be used as base condition for linear regression. While creating dummy variables, drop_first =True, omits first variable and creates dummy variables for other variables

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



Pair-Plots

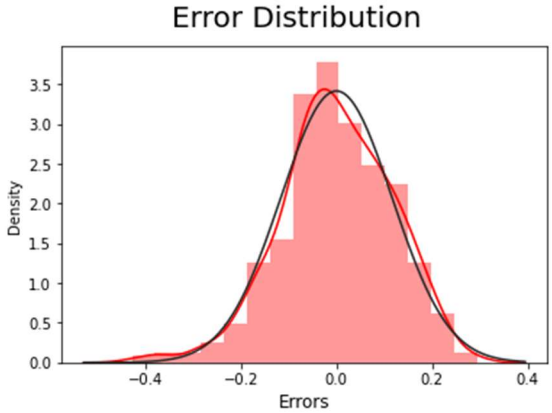
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

➔ i. X & Y are independent.

From above pair plots the predictor variable “temp” and target variable “cnt” has linear relationship. Also we can at least surely say, that variables “hum” & “windspeed” don’t have non-linear relationship with “cnt”. They have less correlation and seem to be linear.

ii. Error terms are normally distributed

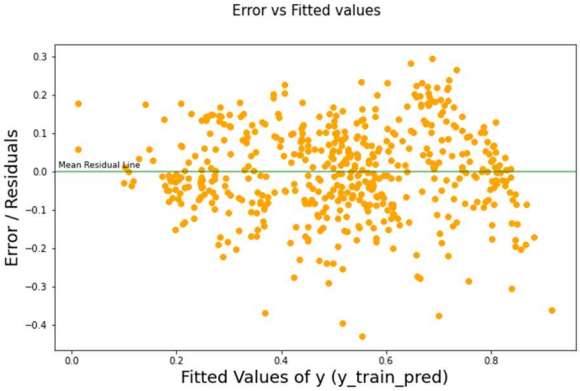
➔ Through residual analysis we can see in plot the Error terms are normally distributed.



iii. Error terms are independent of each other

iv. No Heteroskedasticity

To check iii & iv scatter plot of Residuals or error terms plotted against Fitted y values. And no any time-series pattern is observed. Error terms are independent from each other



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   →Following is the linear model equation for the Bike rental case study.

Cnt = 0.0576* const + 0.2322 * yr +(-0.0399) * holiday + 0.0423 * workingday + 0.5747 * temp + 0.0518 * Monday +

0.0034 * Tuesday + (-0.2369) * Light Snow

The “temperature” has strongest positive impact on bike rentals

Next is the weather conditions have considerable negative corelation with count. During light snow condition rentals would reduce.

Also year has positive correlation on rentals. Basically 2019 sees increase in Bike rentals.

General Subjective Questions

1. Explain the linear regression algorithm in detail

   EDA

   The data received, for any linear regression problem has to be cleaned make fit for regression analysis. In general EDA is done as

   a. Data Understanding
   b. Data cleaning/conversion
      (data type conversions, unnecessary column such as indices or Sr. No. removals, null value more than 70% columns removal)
   c. Data preparation
   d. Univariate analysis
      Each variable shall be looked out thoroughly through boxplots, count plots & overall description such as mean, median, mode & quantiles
   e. Bivariate analysis (if required)
      Correlation between target variable and predictor variable can be plotted for both categorical & continuous variables.
   f. Identify missing values (treatment shall only be done on train set)

   Dummy variables or One hot encoding

   Categorical variables are converted to 0s & 1s format, as regression model would become simpler.

   Test-Train split

   If separate test data is not available then existing data shall be split for learning the coefficients and then validating those on test data. Test-train data split generally done in 70-3-, 60-40 or 80-20. No learning to be done on test data.

   Feature scaling

   Predictor variables or Features can have values of different ranges. All variables of model are brought in same scale by normalize or standardize scaling. This makes interpretation easier and convergence also faster.

   Model fitting

   After scaling, either by using SKlearn or Statsmodel model fit is done. The coefficient and other model parameters detailed summary can be gain with statsmodel liabrary. The summary involves measures such as R2, adjusted R2, F-statistics, P-values of features. These are required to see accuracy of model fit.

   Feature selection

   In case of multiple linear regression, number of features would be available. But to make model simpler, interpretable limited features shall be selected. Feature selection done with technique Recursive feature elimination and variable inflation factor which checks the multicollinearity within features. Multicollinear features could make interpretation difficult, hence generally removed with high collinearity. Along with RFE, VIF, p-value is also check for individual features, which entails the significance of that feature relation with target variable.

   Residual Analysis

Residual or error term analysis done to check assumptions of linear regression are fulfilled or no. Error term is nothing but difference between predicted target variable and actual target variable from train set.

Test data scaling

Test data is scaled as well with scaler.transform function. And not explicitly fit in 0-1 or -1 to 1

Prediction with test data

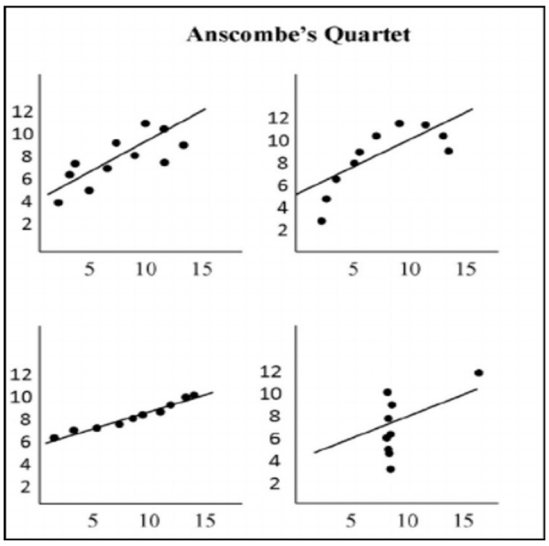
The learned linear regression model (lm) is used to predict target variable for test data. Selected features from test data are fed to model with correlation coefficients to get predicted target test variables.

R2 score check

R2 value checked on test data, shall have range near to R2 value of train data.

2. Explain the Anscombe's quartet in detail.

→ Anscombe's quartet is basically the 4 different data sets with same summary statistics.
In general summary statistics would involve mean, standard error, regression line fit. And these statistics can be same for data with possible outliers or possible data patterns within data set. Hence Anscombe's quartet entails the need to plot the data and visualize and avoiding the inferences based only on descriptive statistics.
Following image shows the 4 data sets plotted, with almost same summary statistics



Source: ResearchGate

3. What is Pearson's R?
→ Pearson's R is a correlation coefficient which measures the linear correlation between two variables. Its value is between -1 to 1.
-1 coefficient -> Strong inverse correlation i.e. increase in X decrease in y
+1 coefficient -> Strong positive correlation i.e. increase in X increase in y
0 -> no relation between X & y
0.05 or -0.05 -> weak relation

4.What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
→Scaling is factorization of the data such that extreme values would be converted to 1, -1 or 0. And all other values then will be between either -1 to 1 or 0 to 1, depending on type of scaling.

Types:

a. Normalized or MinMaxScaler : this scales the value in range 0 to 1
b. Standard scaler : this scales the value in range -1 to 1

Purpose: The output of linear regression is the correlation coefficient. In case of simple linear regression there will be 1 predictor variable & coefficient can explain the variance without any issue. But in case of multiple linear regression there are multiple variables with different value ranges. But correlation coefficient will be from -1 to 1. In this case coefficient values for all variables will be difficult to interpret. The coefficient 0.5 will not mean same for other variable with different range.
Also scaling helps in faster conversion of model.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   →VIF is the correlation between the predictor variables given by 1/(1-R^2). One of the variables is considered as target variable and then R^2 here is how well variance of target variable is explained by other variables

S. No	Mass (g)	Mass(kg)
1	500	0.5
2	300	0.3
3	700	0.7

   In above case same mass is mentioned in both columns but with different units. Hence these both columns are perfectly correlated with coefficient 1. In such scenario R^2 will be 1, if only these 2 variables considered & the VIF for this will be infinite.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   →Q-Q plot or Quantile-Quantile plot is the plot of quantile values within given data. This plot can be compared with normally distributed line or uniformly distributed line. One can take a judgment based on the pattern of Q_Q plot if data distribution normal, uniform or skewed.

   In case of linear regression Q-Q plot can be used to analyse train & test data set. When regression model learns on train data with certain distribution, it would give best results on test data if test data has similar distribution. Else, if distribution of train & test data is not similar, then this factor has to be considered while predicting the target value on test data.