Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

The categorical variables referred were : season and weathersit.

- Here are the key observations which was noticed.

From the plots above we can make the following inferences:

- Demands for bikes is highest in Clear/Partly Cloudy weather, especially in the 'Fall' season
- Demand for bikes is least in the light snow/rain conditions
- Demand for bikes is most during the fall season on average

Q2. Why is it important to use drop_first=True during dummy variable creation?

- The concept of ‘Dummy Variables’ need to be used in order to convert the categorical variables into continuous dummy variables out of these (in order to make Linear Regression model)

- The first variable is redundant and would easily be explained by other dummy variables created for that particular categorical variable. Moreover, since we can have multiple categorical variables in a particular dataset, not using ‘drop_first = True’ would mean that we will have lot of redundant variables, which would make the analysis more complex.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

- ‘temp’ and ‘atemp’ . Both temp and atemp has a correlation coefficient of 0.63 with ‘cnt’ as well.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

I did the Residual Analysis of the Train Data as per below :

- I plotted a histogram of the error terms **(y_train - y_train_cnt)** using a distplot

- After plotting the distplot of error terms, I validated if the error terms were normally distributed and it was exactly the same in my case
- Linear Regression Model Acceptance Criteria

  - The model does not overfit. - The R-Squared & Adjusted R-squared values should be very close to each other(less than 0.5% difference) on both training and test datasets. - The model is simple enough to be understood - The model is built using significatn features. - The final model's test and predictions should have R-squared values closer to the training set. - The residuals are normally distributed

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

As per the final model which I have built, below are the top 3 features significantly towards explaining the demand of the shared bikes.

- ‘temp’
- Weathersit
- weekday

General Subjective questions

Q1. Explain the linear regression algorithm in detail.

Regression is a method of modelling a target value based on independent predictors. This method is mostly used for forecasting and finding out cause and effect relationship between variables. Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.

- Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(X) and dependent(y) variable.
- For Simple linear regression, an example of a model in terms of y and x equation is , y = b0+b1X

- Multilinear regression is a type of regression analysis where the number of independent variables will be greater than 1 and there is a linear relationship between the independent variables (X) and dependent(y) variable.
- For Multi linear regression, an example of a model in terms of y and X0,X1,X2…Xn equation is , y = b0+b1X1+b2X2+….+bnXn

Q1. cont..

- Considering Simple Linear Algorithm for explanation here, the motive of the linear regression algorithm is to find the best values for b0 and b1 so as get the best fit line
- **Cost Function -** The cost function helps us to figure out the best possible values for b0 and b1 which would provide the best fit line for the data points. Since we want the best values for b0 and b1, we convert this search problem into a minimization problem where we would like to minimize the error between the predicted value and the actual value.

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

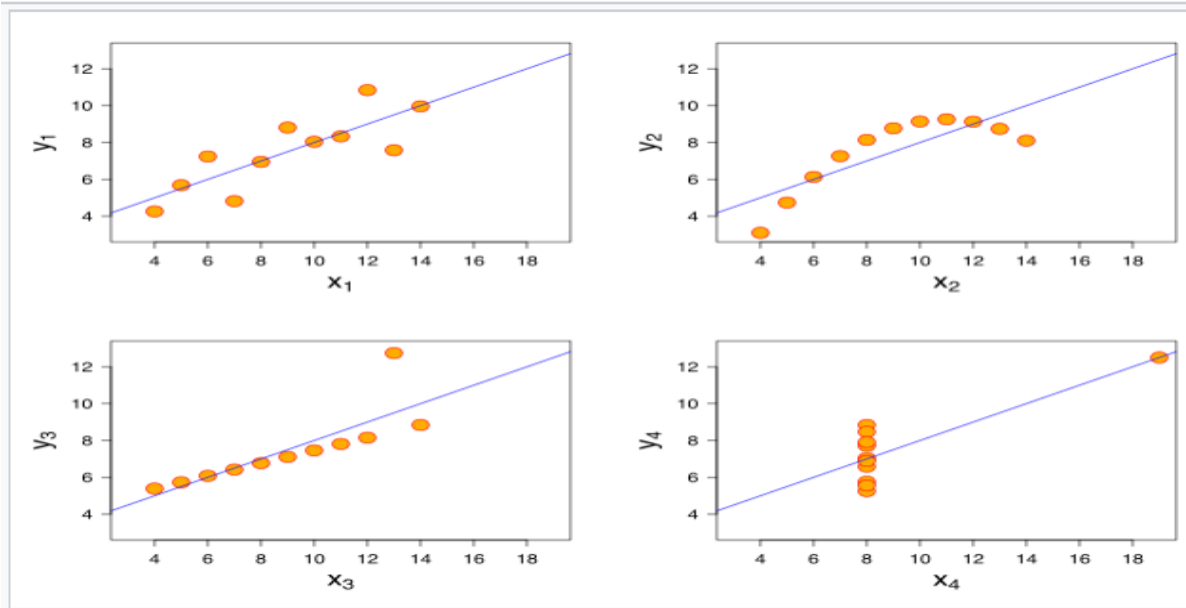
$$J = \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

Q1. cont..

- The difference between the predicted values and ground truth measures the error difference
- We square the error difference and sum over all data points and divide that value by the total number of data points. This provides the average squared error over all the data points. Therefore, this cost function is also known as the Mean Squared Error(MSE) function.
- The coding part of getting the best fit line involves the below concepts
  - Reading the dataset and understanding it
  - Finding continuos and categorical variables in the dataset
  - Making appropriate changes to the dataset wherever required
  - Visualization of Original Dataset including performing EDA to understand various variables & checking the correlation between the variables
  - Data Preparation which involves creation of dummy variables for all the categorical variables
  - Dividing the data to train and test
  - Perform Scaling
  - Divide the data into X and y
  - Data Modelling and Evaluation which involves creating Linear regression model using automated(RFE) approach or Manual approach or a mixed approach, checking various assumptions & checking the Adj R-square for both Train and test data & reporting Final Model.

Q2. Explain the Anscombe’s quartet in detail.

- Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and 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 and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

Q2. cont....

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.
- The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.]

Q2. cont....

- The datasets are as follows. The x values are the same for the first three datasets.

Anscombe's quartet

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

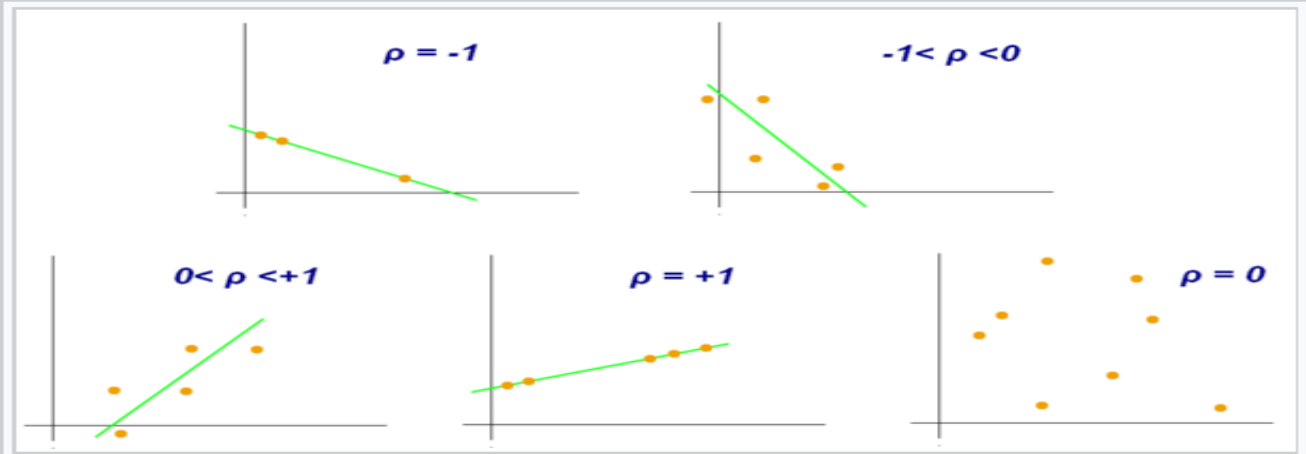
Q2. cont....

- For all 4 datasets :

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : σ^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : σ^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

Q3. What is Pearson’s R?

- In statistics, the Pearson correlation coefficient (PCC, pronounced /ˈpɪərsən/), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC) or the bivariate correlation,[1] is a measure of the linear correlation between two variables X and Y. According to the Cauchy–Schwarz inequality it has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation. It is widely used in the sciences.



Examples of scatter diagrams with different values of correlation coefficient (ρ)

Q3. contd…

- **Assumptions**
- The For the Pearson r correlation, both variables should be normally distributed. i.e the normal distribution describes how the values of a variable are distributed. This is sometimes called the ‘Bell Curve’ or the ‘Gaussian Curve’. A simple way to do this is to determine the normality of each variable separately using the Shapiro-Wilk Test.
- There should be no significant outliers. We all know what outliers are but we don’t know the effect of outliers on Pearson’s correlation coefficient, r. Pearson’s correlation coefficient, r, is very sensitive to outliers, which can have a very large effect on the line of best fit and the Pearson correlation coefficient. This means — including outliers in your analysis can lead to misleading results.
- Each variable should be continuous i.e. interval or ratios for example weight, time, height, age etc. If one or both of the variables are ordinal in measurement, then a Spearman correlation could be conducted instead.

Q3. contd…

- **<u>Assumptions</u>**
- The two variables have a linear relationship. Scatter plots will help you tell whether the variables have a linear relationship. If the data points have a straight line (and not a curve), then the data satisfies the linearity assumption. If the data you have is not linearly related you might have to run a non-parametric
- The observations are paired observations. That is, for every observation of the independent variable, there must be a corresponding If there are 12 observations of weight, you should have 12 observations of age. i.e. no blanks..
- **Homoscedascity.** Homoscedascity simply refers to ‘equal variances’. A scatter-plot makes it easy to check for this. If the points lie equally on both sides of the line of best fit, then the data is homoscedastic. As a bonus — the opposite of homoscedascity is heteroscedascity which refers to refers to the circumstance in which the variability of a variable is unequal across the range of values of a second variable that predicts it

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

- Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

Techniques to perform Feature Scaling

- **Min-Max Normalization:** This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

Q4. cntd…

- **Standardization:** It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

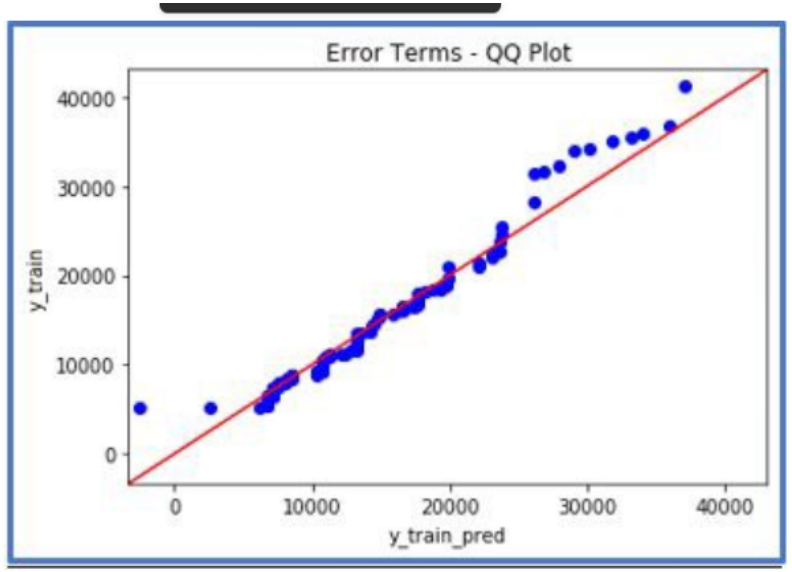
$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

- VIF = Variance Inflation Factor
- In VIF, each feature is regression against all other features. If R2 is more which means this feature is correlated with other features.
- VIF = 1 / (1 – R2)
- When R2 reaches 1, VIF reaches infinity
- An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

Q6. contd...

- **X-values < Y-values:** If x-quantiles are lower than the y-quantiles. q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.



- Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis

Q6. contd...

- **Python Usage :** statsmodels.api provide qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively.
- **X-values < Y-values:** If x-quantiles are lower than the y-quantiles. q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Advantages :

- It can be used with sample sizes also
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

Scenarios where it can be used:

If two data sets-

- come from populations with a common distribution
- have common location and scale
- have similar distributional shapes
- have similar tail behavior