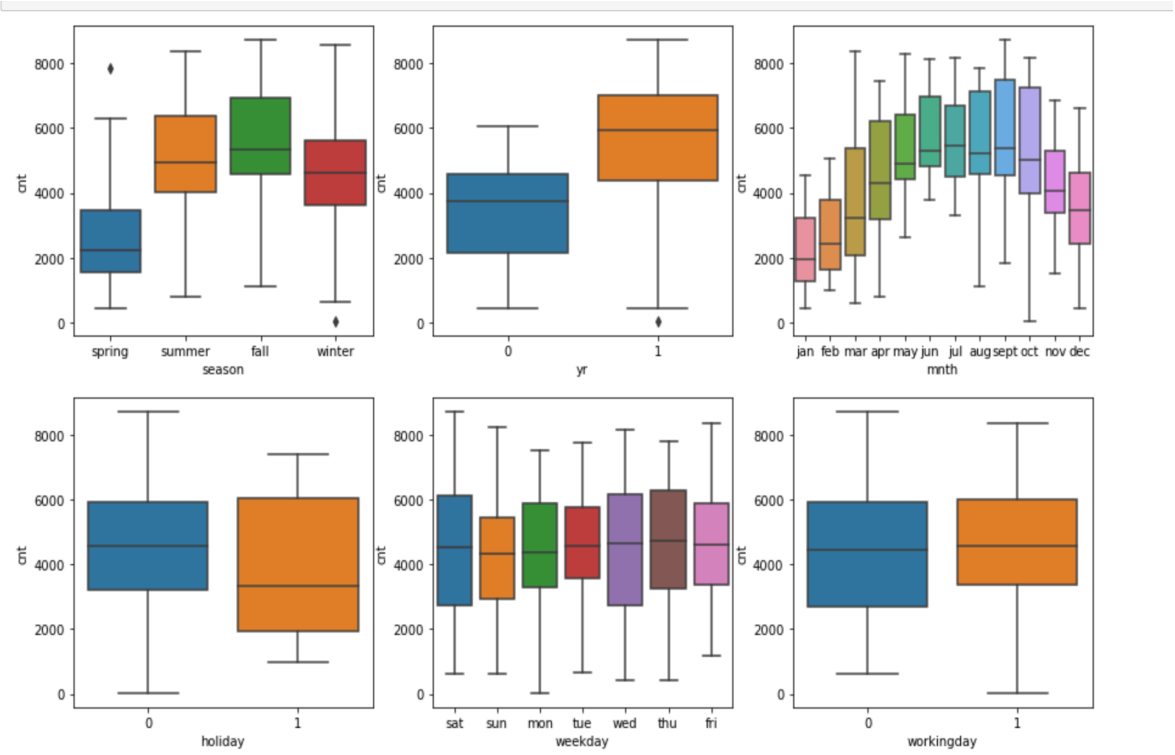
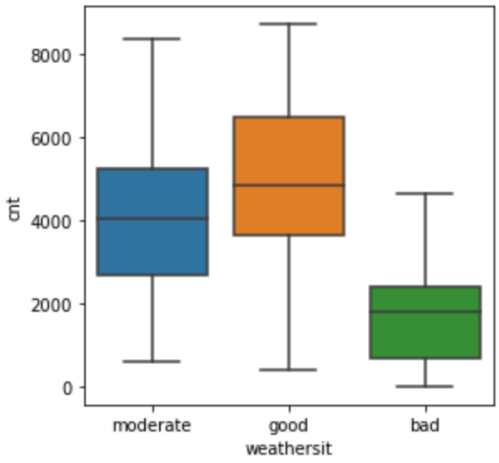
Table of Contents

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)

Observations from above boxplots for categorical variables:
The year box plots indicates that more bikes are rent during 2019.
The season box plots indicates that more bikes are rent during fall season.
The working day and holiday box plots indicate that more bikes are rent during normal working days than on weekends or holidays.
The month box plots indicates that more bikes are rent during september month.
The weekday box plots indicates that more bikes are rent during saturday.
The weathersit box plots indicates that more bikes are rent during good weather



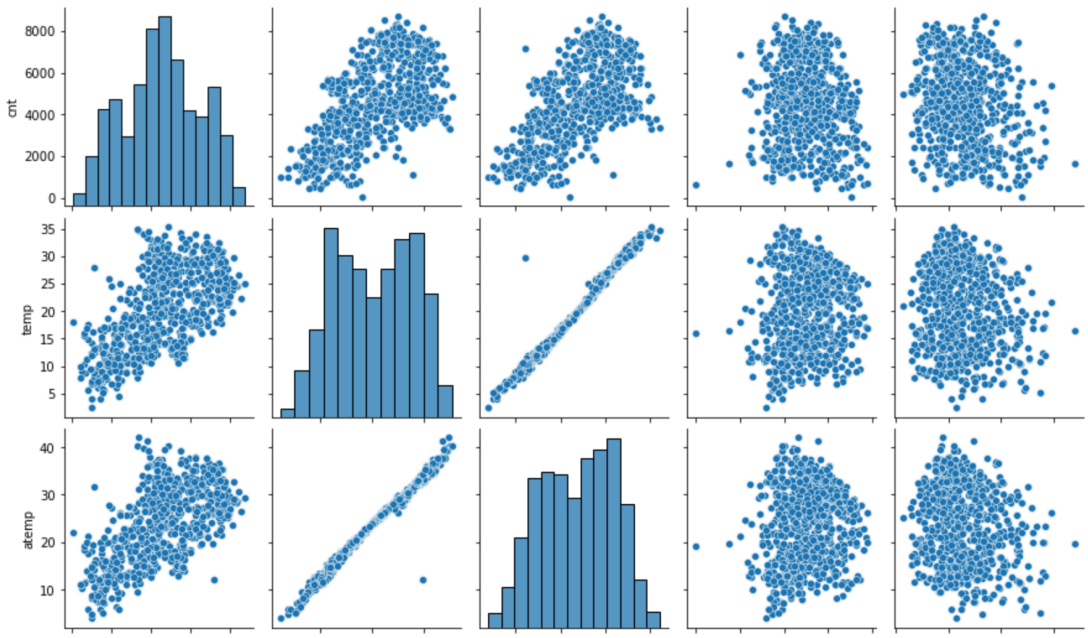


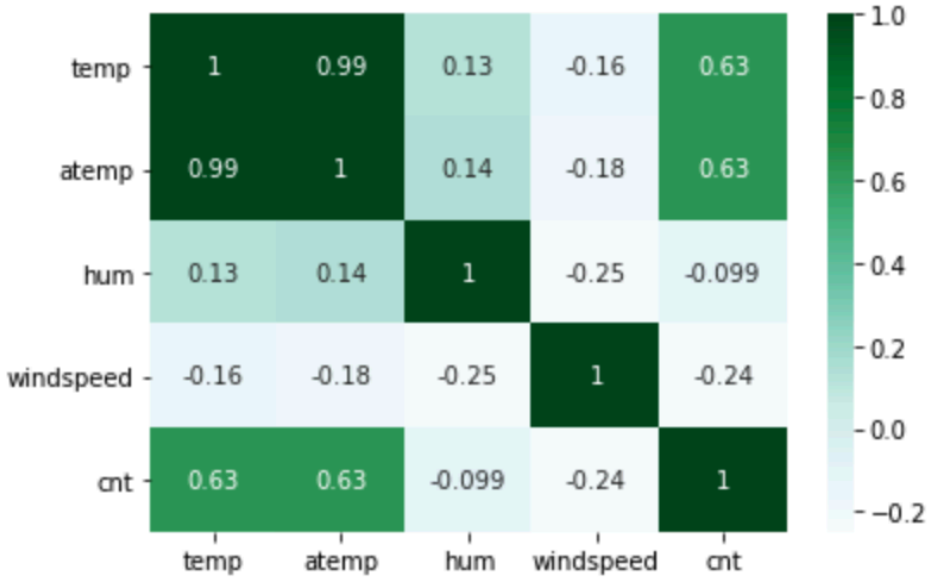
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished.

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

By looking at the pair plot temp and atemp variable has the highest (0.63) correlation with target variable 'cnt'.





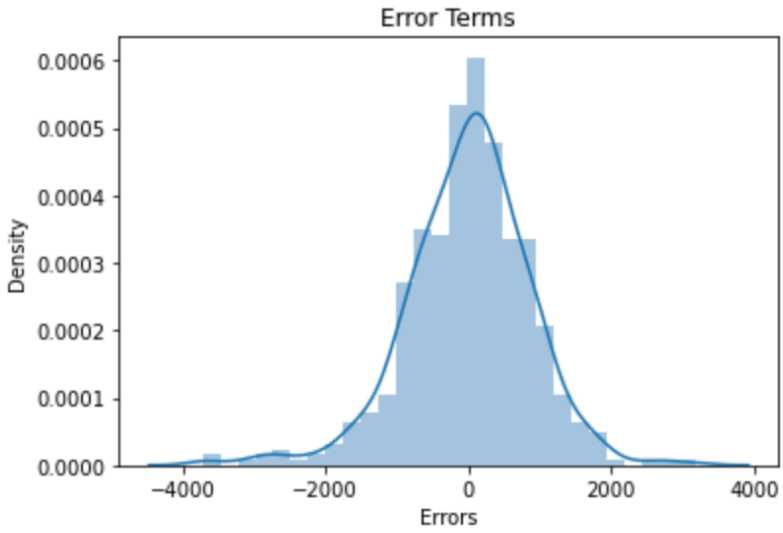
images for visualizing categorical variables

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

important assumptions in regression analysis:

There should be a linear and additive relationship between dependent (response) variable and independent (predictor) variable(s). A linear relationship suggests that a change in response Y due to one unit change in X^1 is constant, regardless of the value of X^1. An additive relationship suggests that the effect of X^1 on Y is independent of other variables.
There should be no correlation between the residual (error) terms.
The independent variables should not be correlated. Absence of this phenomenon is known as multicollinearity.
The error terms must have constant variance.
The error terms must be normally distributed.

```
: plot_res_dist(y_train, y_train_pred)
```



Errors are normally distribured here with mean 0. So everything seems to be fine¶

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

the Top 3 features contributing significantly towards the demands of share bikes are:
weathersit_Light_Snow(negative correlation).
yr_2019(Positive correlation).
weekday(positive)
temp(positive)
temp(Positive correlation).

General Subjective Questions

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

Regression analysis is a statistical technique for investigating and modeling the relationship between variables.In linear regression model x is called the independent variable and y is called the dependent variable. two types of linear regression models are there.
If it involves only one dependent variable, it is called **a simple linear regression model.** In general, the response variable y may be related to k variables, x_1, x_2,…, x_k, so that is called a **multiple linear regression model.**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Linear Regression is part of model building and model evaluation using the clean and prepared data build the model.

Simple linear regression algorithm Steps:

- read and visualise the dataset using seaborn
    - Import data using the pandas library
    - understanding the structure of the data
    - Visualising the data using seaborn, first make a pairplot of all variables present to visualize which variables are most correlated to the dependent variable .
    - Then perform the simple linear regression using the most correlated feature variable.
- Performing Simple Linear Regression
    - Assign feature variable a value X and the response variable to y
    - split the dataset into train and test sets by importing train_test_split from the sklearn.model_selection library. Good practice is to keep 70% of data in train dataset and 30% in the test dataset.
- build the model on the training data and draw inferences.ie .make a linear model using two different libraries: statsmodels\ SKLearn.

$$Y = \beta_0 + \beta_1 X + \epsilon$$

    - **I**mport the statsmodel.api library for performing linear regression
    - manually use add_constant attribute and add constant to X_train dataset
    - fit a regression line using OLS Ordinary least squares
    - check the summary and find the key statistics like p values, R-squared and the F statistic
    - **plot using scatter plot**
- Residual analysis to be done for validating the assumptions of the model, the error terms are normally distributed or not by using distplot and look for the patterns in residuals and no pattern should be there for a good model.
- Do Predictions on the test data set :
    - Same like train data set add a constant to X_test and use predict attribute to predict y values.
    - check r-squared on the test set
    - Visualizing the fit on the test set

The same can be done using sklearn also

Multiple linear regression algorithm Steps:

- read and visualise the data
    - Import pandas and numpy library
    - Read the data
    - understanding the structure of the data
    - Visualising the data and find if any multicollinearity also identify predictors which has strong association with outcome variable.
    - visualise the numeric variable by pairplot
    - visualize the categorical variable by boxplot

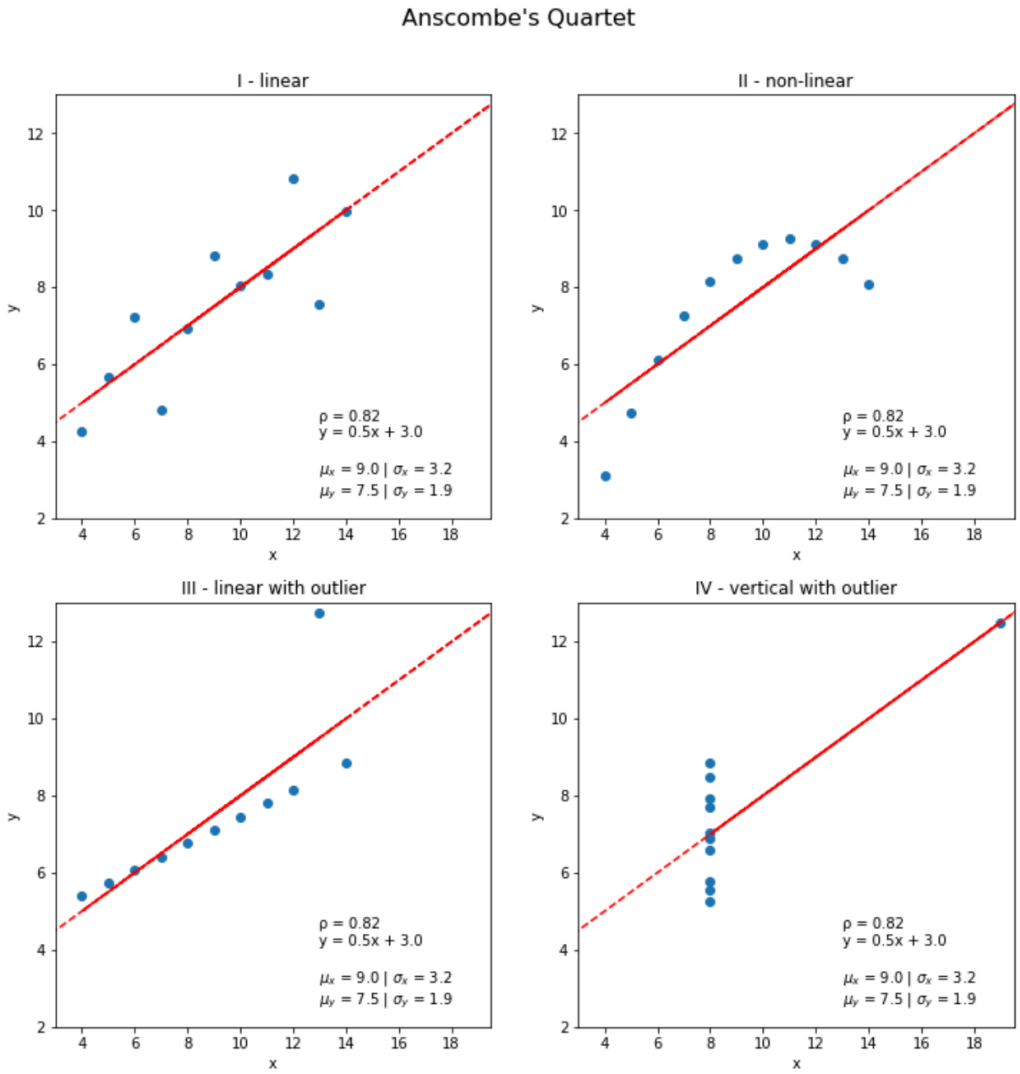
- Data Preparation
    - convert any yes to 1 and no to 0 if it has any
    - Create a dummy variable for the categorical variable
    - after creating dummy variables drop the unnecessary double columns

- Split the data into training and testing sets
    - rescale the features by min-max scaling or standardisation
    - split the data into training and testing set
- Build a linear model using statsmodels
    - add constant
    - create a fit model
    - visualize the data with a scatter plot and the fitted regression line
    - we can do the same by adding one variable at a time and doing the same steps
    - or We can add all the variables to the model at starting and we can remove the variables after checking the highest p value and after checking the VIF value, VIF less than 5 is good
    - Drop the variables and update the model, check the VIF again and do the same , high VIF and high p values variables should be dropped
- Residual analysis of the train data
    - error terms should be normally distributed
- make predictions on the data
    - apply scaling on the test set
    - add constant variable
    - make predictions using model
- model evaluation
    - plot graph for actual vs predicted values

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

Anscombe's quartet is a classic example that illustrates why visualizing data is important. The quartet consists of four datasets with similar statistical properties. Each dataset has a series of x values and dependent y values 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.

This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc



Summary statistics are very helpful when we're getting to know the data, but be wary of relying exclusively on them. Remember, statistics can mislead; be sure to also plot the data before drawing any conclusions or proceeding with the analysis

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

A linear relationship exists between two quantitative variables when there is an overall tendency for increases in the value of one variable to be accompanied by increases in the other variable (a positive relationship), or for increases in the first to be accompanied by decreases in the second (a negative relationship) The Pearson product moment correlation coefficient, r, is a measure of the degree of relationship between two variables, x and Y, based on the discrepancies of the subjects' paired z scores, zX – zY. r varies between –1 and +1, which represent perfect negative and perfect positive linear relationships, respectively

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation

coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

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

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. . If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling t is important to note that scaling just affects the coefficients

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ). it loses some information in the data, especially about outliers.

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

sklearn.preprocessing.scale helps to implement standardization scaling in python

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

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables. For example, we would fit the following models to estimate the coefficient of determination R1 and use this value to estimate the VIF:If there is perfect correlation, then VIF = infinity. This

shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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).

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. 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. 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). 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. A general rule of thumb is that if VIF > 10 then there is multicollinearity.

If VIF is large and multicollinearity affects your analysis results, then you need to take some corrective actions before you can use multiple regression. Here are the various options: One approach is to review your independent variables and eliminate terms that are duplicates or not adding value to explain the variation in the model.

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

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.