Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Ans 1. Inferences from the analysis are as follows:

- Significant predictors for the demand of shared bikes are:
    - year,
    - holiday,
    - temperature,
    - windspeed,
    - weathersit(Light, Mist),
    - season(summer, winter),
    - Month(September)

- Most significant predictor is temperature followed by Year & Winter season
- Holiday, Windspeed & weathersit (Light, Mist) predictors have negative effect on the demand.
- As Year is among top 3 predictors, it’s quite evident that compared to year 2018, there is a remarkable growth in the demand of shared bikes.
- Company can increase the supply in summer and winter season are target is positively related to these seasons as compared to other seasons.
- Company can decrease the supply on holidays as people do not prefer bike sharing on these days which can be understood by negatively related predictor to target.
- Company should prioritize its supply especially during September month.

Q2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

Ans 2. Drop_first=True is useful for reducing the number of extra columns created during the creation of dummy variables. In this way, correlations between dummy variables are reduced. For instance, if we have 3 types of values in the Categorical column and we wish to create a dummy variable for that column. It is obvious that a variable is unfurnished when it is not furnished and semi-furnished. It is therefore unnecessary to include a third variable to identify unfurnished units.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

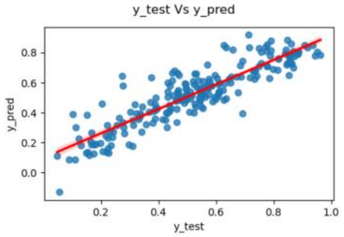
Ans 3. Among the numerical variables, **Temperature** variable has the highest correlation with the target variable.

----

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

Ans 4. I have validated the assumptions of Linear Regression Model based on below points:

Linear Relationship: The relationship between the predicted and actual values was linear in both training and test dataset.



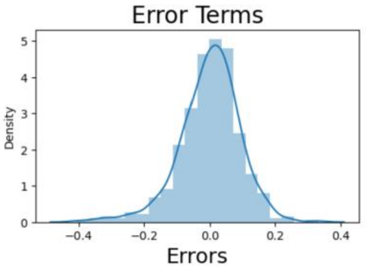
No Autocorrelation: The linear regression analysis requires that there is little or no autocorrelation in the data. Autocorrelation occurs when the residuals are not independent of each other.

Durbin-Watson (DW) Test is generally used to check the Autocorrelation. Range of Durbin Watson Test for the 7th final model was 2.097 i.e. Negative Autocorrelation.

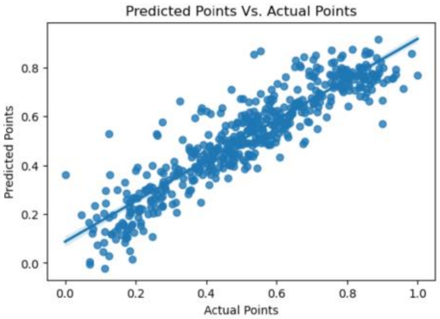
Mostly the values are from 0 to 4, where

- 0-2 shows positive Autocorrelation
- 2 means No Autocorrelation and
- 2-4 means Negative Autocorrelation.

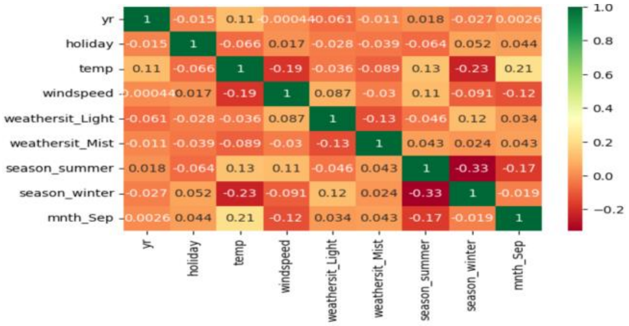
Multivariate Normality: The linear regression analysis requires all variables to be multivariate normal i.e. Means data should be normally distributed. In the final model, Error terms were normally distributed with mean centered at zero:

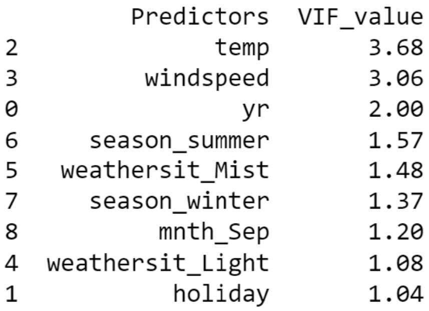


Homoscedasticity: Homoscedasticity describes a situation in which the error term is the same across all values of the independent variables. There was no visible pattern in the scatter plot of residual values vs predicted values, which is a good way to check for homoscedasticity.



No or low Multicollinearity: In the heat map all correlation coefficients for all pairs of predictor variables were very less i.e. <0.21. Also VIF for all predictors are <5. Therefore, it can be concluded that there was no multicollinearity in the model.



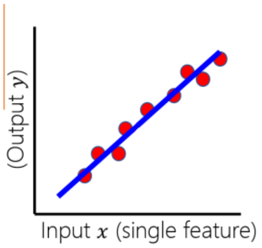

```
          Predictors  VIF_value
2               temp       3.68
3          windspeed       3.06
0                 yr       2.00
6      season_summer       1.57
5     weathersit_Mist       1.48
7      season_winter       1.37
8           mnth_Sep       1.20
4   weathersit_Light       1.08
1            holiday       1.04
```

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
Ans 5. Top 3 features contributing significantly towards explaining the demand of the shared bikes are: **temperature, Year & Winter season**

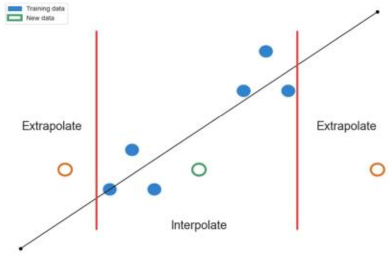
****

General Subjective Questions
Q1. Explain the linear regression algorithm in detail. (4 marks)

Ans 1. An analysis of linear relationships between independent variables and dependent variables is known as linear regression. It is widely used for Forecasting and Prediction. Linear regression is a form of parametric (using a finite number of parameters) regression.



Regression guarantees ‘interpolation’ but not necessarily ‘extrapolation’. Interpolation means using the model to predict the value of a dependent variable on the independent values that lie within the range of data you already have. Extrapolation, on the other hand, means predicting the dependent variable on the independent values that lie outside the range of the data the model was built on.



Linear regression types:

- Simple linear regression: used when the number of independent variables is 1.

  The relationship can be represented by the following equation Y = mX + c + e

  - Y is the dependent variable we are attempting to predict.
  - X is the independent variable we use to make predictions.
  - m is the slope of the line, representing the change in y for a unit change in x.
  - b is the y-intercept, the value of y when x is zero.
  - E is the error variable

- Multiple linear regression: used when the number of independent variables is more than 1.

  $$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \epsilon$$

  - y = the predicted value of the dependent variable
  - B_0 = the y-intercept (value of y when all other parameters are set to 0)
  - ϵ = model error

Aspects to consider when moving from simple to multiple linear regression are as follows:

1. Overfitting:

   - The model may become far too complex by adding variables
   - It may end up memorizing the training data and, consequently, fail to generalize.
   - A model is generally said to over fit when the training accuracy is high while the test accuracy is very low.

2. Multicollinearity:
3. Feature selection: Selecting an optimal set from a pool of given features.

   - Manual feature selection
   - Automated feature selection

- Finding a balance between the two

Assumption of Linear Regression:

- There is a linear relationship between X and Y.
- Error terms are normally distributed with mean zero.
- Error terms are independent of each other.
- There is very little or no auto-correlation in the data.
- Error terms have constant variance (homoscedasticity).

Linear regression finds the best-fitting line to model the relationship between variables. It involves defining a linear equation, determining the error using a cost function, and optimizing the parameters through gradient descent during the training phase. The trained model can then be used for making predictions.

The equation of the best fit regression line $Y = \beta_0 + \beta_1X$ can be found by minimizing the cost function, which is done using the following two methods:

- Differentiation
- Gradient descent

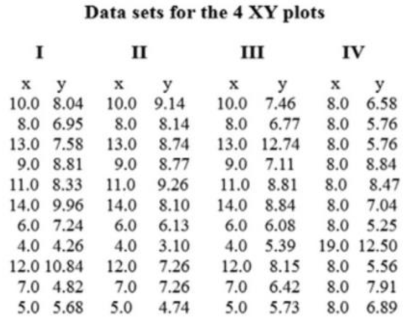
The strength of a linear regression model is mainly explained by R^2, where $R^2 = 1 - (RSS/TSS)$.

- RSS: Residual sum of squares
- TSS: Total sum of squares

Q2. Explain the Anscombe's quartet in detail. (3 marks)

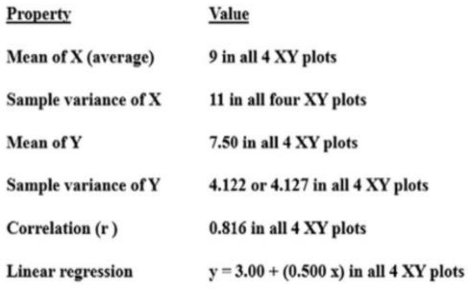
Ans 2. Anscombe's quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph. The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data.



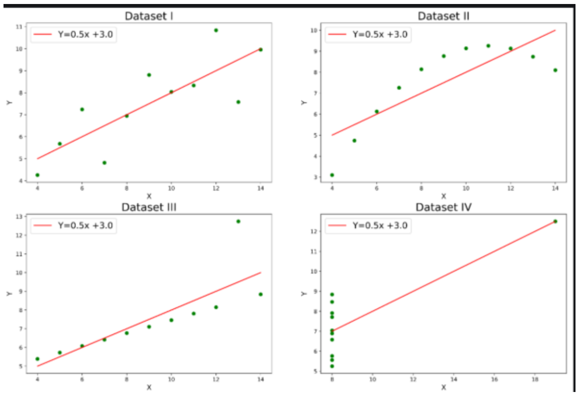
Data sets for the 4 XY plots

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	5.76
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	8.81	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	7.26	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



Property	Value
Mean of X (average)	9 in all 4 XY plots
Sample variance of X	11 in all four XY plots
Mean of Y	7.50 in all 4 XY plots
Sample variance of Y	4.122 or 4.127 in all 4 XY plots
Correlation (r)	0.816 in all 4 XY plots
Linear regression	y = 3.00 + (0.500 x) in all 4 XY plots

When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.



Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

Q3. What is Pearson's R? (3 marks)

Ans 3. The Pearson correlation coefficient, also known as Pearson's r, measures the strength of a relationship between two variables and their association with one another. Pearson's Correlation Coefficient is named after Karl Pearson. Only a linear relationship between two continuous variables can be tested by the Pearson correlation

Pearson's Correlation Coefficient Formula:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

Where

r = Coefficient of correlation

xbar = Mean of x-variable

ybar = Mean of y-variable.

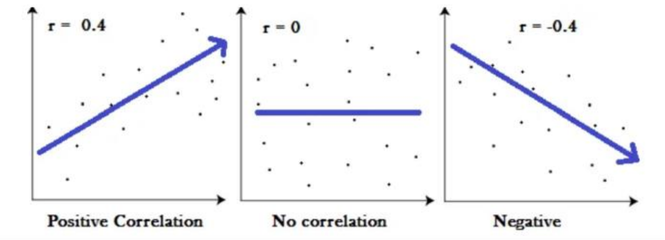
xi yi = Samples of variable x, y

The value of correlation coefficients lies between -1 and +1.

0 indicates no correlation.

1 indicates a perfect positive correlation.

-1 indicates a perfect negative correlation.



The magnitude indicates the relationship's strength, while the sign indicates the relationship's direction.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans 4.

What is scaling?

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.

Why is scaling performed?

Scaling is done for a number of purposes:

- Scaling guarantees normalization i.e. all features are on a comparable scale and ranges.
- When scaled, algorithms perform better or converge more quickly.
- Numerical instability can be prevented by avoiding significant scale disparities between features.
- Without scaling, bigger scale features could dominate the learning, producing skewed outcomes.

What is the difference between normalized scaling and standardized scaling?

Range:

- Normalized scaling confines the data to a specific range, usually between 0 and 1.
- Standardized scaling transforms the data to have a mean of 0 and a standard deviation of 1.

Sensitivity to Outliers:

- Normalized scaling is sensitive to outliers since it considers the minimum and maximum values in the dataset.
- Standardized scaling is less sensitive to outliers as it relies on the mean and standard deviation, which are less affected by extreme values.

Interpretability:

- Normalized scaling is more intuitive when you want your data to be within a certain range.

- Standardized scaling is preferable when you want to compare the relative positions of data points across different features.

Formula:

- The formula for normalized scaling is:

$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here,

X is the original data point,
Xmin is the minimum value in the dataset, and
Xmax is the maximum value.

- The formula for standardized scaling is: $z = \frac{x - \mu}{\sigma}$

Here,

X is the original data point,
μ is the mean of the dataset, and
σ is the standard deviation.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans 5.

A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model.

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates a perfect correlation between two independent variables. In the case of perfect correlation, we get R-square =1 & VIF=1/ (1-Rsquare) i.e. infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans 6.

A Quantile-Quantile (Q-Q) plot is a plot of the quantiles of the first data set against the quantiles of the second data set. The pattern of points in the plot is used to compare the two distributions. The main step in constructing a Q–Q plot is calculating or estimating the quantiles to be plotted.

It is used to check following scenarios:

If two data sets —

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

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. The presence of outliers can also be detected from this plot.
