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)***

   Answer: From the categorical variables we see the following patters on the target variable cnt

   - Demand of bikes were more in 2019 than in year 2018
   - Business operated in all months of year and all seasons
   - Demand is more when weather is- weathersit =:
     1: Clear, Few clouds, Partly cloudy, Partly cloudy
     2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
   - There is no demand in case the weather is 4 (Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog)
   - Average(median) demand of bikes was lesser in the spring season compared to other seasons
   - Demand is higher in the months 3,9 and 10

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

   Answer:

   - ***drop_first=True*** is important to use as during dummy variable creation extra variable is produced which can be easily deduced from other variables created hence to save space we can drop this extra variable
   - (for eg. In case of traffic light if green and red is there as variables we can identify yellow for data points we don't have green and red)

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

   Answer:

   - **atemp** and **temp** has the highest correlation with the target variable cnt

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

   Answer:
   Validation points:

   - Plotted the graph to see if error terms follow a normal distribution with mean 0
   - Plotted graph to see if error terms are independent i.e no visisble pattern in them
   - Plotted graph to see if the error terms have constant variance (**homoscedasticity)**
   - Plotted graph between dependent and independent variables to make sure there was some kind of linear relationship between X and Y

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

Answer : Top three variables are:

- Yr
- Season- spring
- Weather condition - Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

General Subjective Questions

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

Answer:

- **Linear Regression** is a machine learning algorithm based on **supervised learning**. Regression models a target prediction value based on independent variables.
- Different regression models exists based on the kind of relationship between dependent and independent variables, and the number of independent variables getting used.
- Linear regression predicts a dependent variable value (y) based on a given independent variable (x). This regression technique tries to find out a linear relationship between x (input) and y(output).
- While training the model we are given :
  x: input training data (univariate – one input variable(parameter))
  y: labels to data (supervised learning)
- Hypothesis func for this algorithm**- [Y= b0 + b1X]**
- 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 b0 and b1 values.
  b0: intercept
  b1: coefficient of x
- Once we find the best b0 and b1 values, we get the best fit line.
- **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 b0 and b1 values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).
- 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 b0 and b1 values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line the model uses Gradient Descent. The idea is to start with random b0 and b1 values and then iteratively updating the values, reaching minimum cost.

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

Answer:

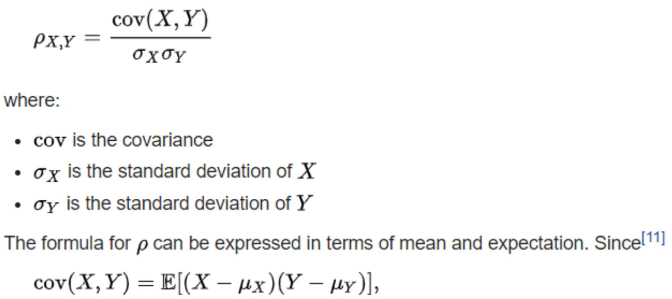
- *Anscombe's quartet* are set of four datasets which have the identical simple statistical property
- yet appear very different when graphed. Each dataset consists of eleven (x,y) points.
- They were constructed in 1973 by the statistician Francis Anscombe

- Used to show the importance of graphing data before analyzing it and the effect of outliers on statistical properties and the inadequacy of basic statistic properties for describing realistic datasets.

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

Answer:

- **Pearson's correlation** (also called Pearson's *R*) is a **correlation coefficient** commonly used in linear regression
- The full name is the **Pearson Product Moment Correlation (PPMC)**. It shows the linear relationship between two sets of data.
- Two letters are used to represent the Pearson correlation: Greek letter rho (ρ) for a population and the letter "r" for a sample.
- Problem with pearson- The PPMC is not able to tell the difference between dependent variables and independent variables.
  Eg- if you are trying to find the correlation between a high calorie diet and diabetes, you might find a high correlation of .8. However, you could also get the same result with the variables switched around. In other words, you could say that diabetes causes a high calorie diet.



$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

The formula for ρ can be expressed in terms of mean and expectation. Since[11]

$$\operatorname{cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)],$$

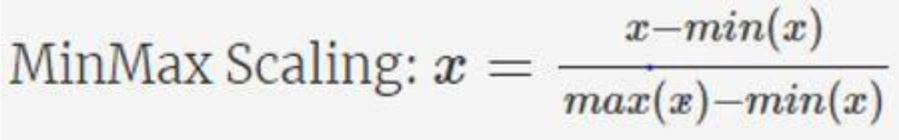
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer:

- scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range.
- It is used in speeding up the calculations in an algorithm.
- If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude

- **Normalization/Min-Max Scaling:**

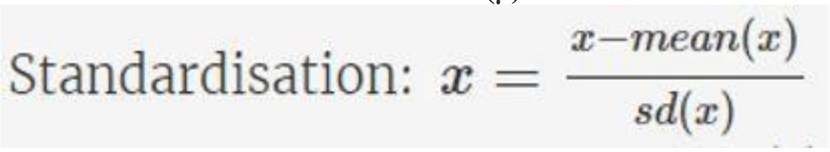
It brings all of the data in the range of 0 and 1. **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.



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

- **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**).



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

- **sklearn.preprocessing.scale** helps to implement standardization in python.

- One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**
   Answer:
   - VIF can be infinite when the case arises if there is a perfect correlation between independent variables, in this case the r2 becomes 1 and putting it in the VIF formula (1/(1-r2)) we will get infinite value
   - In this case we need to drop the variable that causes this perfect relation
   - An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables

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

   Answer:
   - Quantile- quantile plot
   - These are the plots between two quantiles
   - A quantile is a fraction or limit where certain values falls below it
   - Eg. Median is quantile where 50% of data falls below and above it
   - The purpose of Q-Q plot is to see if two sets of data come from 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 sets are linearly related they will lie on same line but not necessarily on the reference line
   - 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.