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 –
   1. People more tend to Rent a bike in fall season followed by summer. In Spring season less people prefer to rent bike.
   2. 2019 is the year where most of the people rented bike compared to year 2018. There is lot difference in median value in both the year.
   3. Seasons and mnth variable are correlated as sales are more in month ranging Aug to Oct (fall season).
   4. There is no significant difference in 75th percentile whether its holiday or not, on other hand More people likes to rent bike over the weekday this may due to they prefer going to offices by renting bike and can be cheaper.
   5. In terms of day of weeks, the median value is approx. same for each day but there are fridays where more people sometimes rent bike sometimes very less people does the same.
   6. The count of renting bikes at 25th, 75th percentile and median are almost identicle, this implies that people tend to ride a bike irrespective of day of week or holiday.
   7. When weather is clear, people opt for ride and rent a bike more. In case of snow falls its very less.

   ---

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   Answer – if column has set as per categories and then we can easily represents data using k-1 (Where k = total categorical representation in column) categorical variables (columns) so there is no need to have additional column for the same representation.
   Take an example of Season column from day.csv dataset (fall, spring, summer, winter), you can represent the same as below

   Eg - 000 will correspond to Fall
   100 will correspond to Spring
   010 will correspond to Summer
   001 will corresponds to Winter

   From above example it is clear that if season is not Spring, Summer or Winter then it is by default Fall (000)

   ---

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

   Answer –

   Considering all the Numerical variables in pair plot, 'registered' variable has highest correlation with target variable 'cnt' followed by 'casual' variables but we can not take these two variables correlation as they are redundant (cnt = casual + registered)

   So we will take '**temp'** variable which has Highest correlation with target variable '**cnt'** (Note – We are not considering 'atemp' variable as it is again redundant as 'temp)

   ---

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

   Answer –

   After plotting Histogram, it is observed that the error terms are normally distributed in range -2 to 2 with mean as zero (0) and they are independent of each other and has constant variance.

   ---

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 –

   The Top3 features contributing significantly towards explaining model are **Temp, yr and season_Winter** and below the equation for bestfit line

   cnt = 0.199434 + 0.490988 X temp + 0.233570 X yr + 0.081741 X seasonWinter + 0.076846 X mnth_Sep + 0.046487 X season_Summer - 0.052057 X mnth_Jul - 0.067169 X season_Spring - 0.080167 X weathersit_Mist - 0.097463 X holiday - 0.147919 X windspeed - 0.284199 x weathersit_Light Snow

   ---

General Subjective Questions

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

Answer –

Linear Regression is linear approach of modelling relationship between dependent and independent variables. This will also state that if the value of one or more independent variable is changed then it will affect dependent variables.

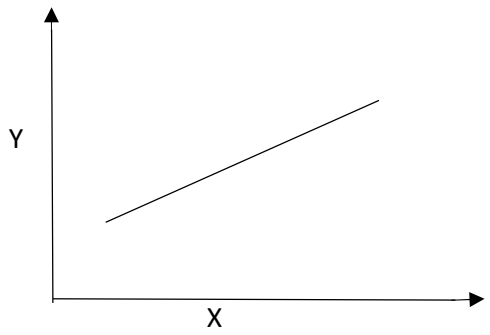
It is represented as

$$y = mX + c \quad OR \quad y = \beta_0 + \beta_1 X$$

where, m= β1 = Slope
c = β0 = Intercept
y = Dependent variable
X = Independent Variable

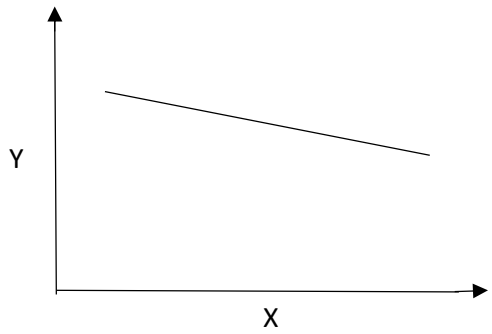
Positive Linear Relationship –

With increase in Independent variables, value of dependent variable also gets increased



Negative Linear Relationship –

With increase in Independent variables, value of dependent variable gets decreased



There are 2 types of Linear Regression

a. Simple Linear Regression
b. Multiple Linear Regression

a. Simple Linear Regression → The Model is built with only 1 Independent variable. The equation mentioned above describes Simple linear regression

$$y = \beta_0 + \beta_1 X$$

b. Multiple Linear Regression → The Model is built with more than 1 Independent variable. It represents the relationship between 2 or more independent variables. The equation mentioned above describes Simple linear regression

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \ldots\ldots + \beta_n X_n$$

where β_1 to β_n = coefficient of independent variables
X_1 to X_n = Independent variables
$\beta 0$ = Constant

sometime we use error (€) in the equation to complete the same.
Coefficient presents, if change in 1 unit of coefficients will change target variable(dependent variable) with 1 unit

Another aspect of Linear regression is **Residuals** which is calculated using

Residuals = measured value – predicted value

$$e_i = y_i - y_{pred}$$

RSS (Residual sum of square) $= e_1^2 + e_2^2 + \ldots + e_n^2$

As we know, $y = \beta_0 + \beta_1 X$

Then,

$$ei = y_i - \beta_0 - \beta_1 X_i$$

Therefore,

$$RSS = (y_1 - \beta_0 - \beta_1 X_1)^2 + (y_2 - \beta_0 - \beta_2 X_2)^2 + \ldots\ldots + (y_n - \beta_0 - \beta_n X_n)^2$$

$$RSS = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 X_i)^2$$

TSS (Total sum of square) - $(y_1 - y_{bar})^2 + ...+ (y_n - y_{bar})^2$

$$TSS = \sum_{i=1}^{n} (y_i - y_{bar})^2$$

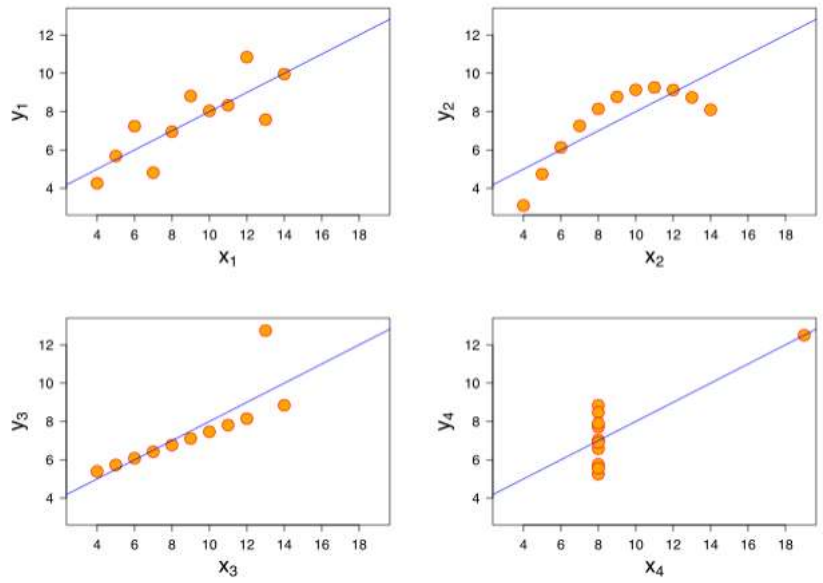
$$R2 = 1 - (RSS / TSS)$$

---

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

Answer –

Anscombe's quartet (constructed in 1973 by Francis Anscombe's) is basically states that if there are 4 simple data sets with 11 pairs of X and Y points and which has close to identical Summary statistics like (mean of X, mean of y, sample variance of X, sample variance of Y, correlation, LR line, coefficient), but when you actually represent as 4 graphs, each graph will tell different story



Out of 4 graphs,

a. first graph will tell Simple Linear relationship with some variance
b. second graph will tell Non-Linear relationship
c. third graph the Distribution is Linear with 1 outlier

d. fourth graph has 1 outlier is enough to produce high correlation coefficient even though other datapoints do not indicate any relationship at all.

With only viewing of statical summary may not be sufficient to check for the data set that's why we need to visualize it for better understanding of relation.
So, to demonstrate both the importance of graphing data before analysing it and the effect of outliers and other influential observations on statistical properties.
(note – graph reference – Wikipedia)

---

3. What is Pearson's R? (3 marks)
Answer –
Pearson's R also known as Pearson's correlation coefficient is a value that measures Linear correlation between 2 variables. It lies between -1 to 1.

a. If the value is -1 then there is Negative correlation between 2 variables
b. If the value is 0 then there is NO correlation between 2 variables
c. If the value is 1 then there is Positive correlation between 2 variables

It also represented as,
Pearson's correlation coefficient is covariance of two variables divided by product of their standard deviation.

$$\rho_{X,Y} = cov(X,Y) / \sigma_X \sigma_Y$$ ---- When applied to Population

where,
cov – covariance
σ_X – Standard deviation of X
σ_Y – Standard deviation of Y

---

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

Scaling – Scaling is the method used for data pre-processing to standardize the feature of dataset to a certain fixed range

Scaling is performed because, numerical variables may not be in same unit due to which machine learning algorithms don't perform well since variables have different scales. So to standardize the values of all variables(features) scaling is performed.
If we don't perform scaling then, machine learning algorithm may give high weightage to greater values and low weightage to smaller values.

Difference between normalized scaling and standardized scaling –

a. Normalized scaling also known as Min Max Scalar, this will rescale all the values of variables between 0 and 1 whereas standard scalar does Not binds values to any fixed range, all the data are normally distributed with mean = 0 and standard deviation = 1
b. Min Max Scaling represented by

   x = (x -min(x))/(max(x) – min(x))

c. Standardized scaling represented by

   x = (x -mean(x))/sd(x) --------- sd = standard deviation

---

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

Answer –

VIF stands for Variation Inflation Factor, this is used to find how well one independent variable is performed by all the other independent variables as combined.

You can measure multicollinearity using VIF, High VIF indicates associate independent variable is highly collinear with other variables in model, if VIF is low then it is not highly collinear with other variable.

VIF is represented by,

$$VIF_i = 1 / (1 – R_i^2)$$

VIF is infinite only when we have R^2 value = 1, which leads denominator as 0 and anything divided by 0 is infinite.

Generally,

If VIF > 10 → High VIF and variable should be removed

If VIF > 5 → This can be OK but its better to Inspect the Variable

If VIF < 5 → This says a good value and we can retain this.

---

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

Answer –

Q-Q plot also known as Quantile-Quantile plot is a plot of two quantiles between each other. The use of this plot is to assess if a set of data comes from some theoretical distribution i.e. Normal distribution or exponential distribution.

While building a linear regression model, we assume that the residuals are normally distributed, we can use a Normal Q-Q plot to check that assumption.
Q-Q plot also allows us 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.
