Assignment-based Subjective Questions and Answers

1. 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. Registrations are increased from 2018 to 2019.

2. Registrations lowest in Jan and increased to maximum between June to Sept further reduced gradually to Dec.

3. During non-holidays registrations are more.

4. No significant variation in day wise registrations.

5. No significant variation between weekday and weekend or holiday registrations.

6. Registrations are lowest in spring and increases in summer and become highest in rainy and reduces in winter.

7. Registrations are lowest in light rain or snow or Thunderstorm with partial clouds increased to Mist weather and highest in Clear weather.

8. Significant variables to consider yr, mnth, holiday, weathersit and season.

9. Out liers are not significant.

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

Ans. When two or more dummy variables created by encoding they are highly correlated (multi-collinear). This means that one variable can be predicted from the others which is difficult to interpret predicted coefficient variables in regression models. Hence to reduce the effect of multicollinearity drop_first=True used to drop the first dummy variable.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation?

with the target variable? (1 mark)

Ans. 'temp ' and 'atemp' has the highest correlation.

4. How did you validate the assumptions of Linear Regression after building the model on the

training set? (3 marks)

Ans. Residual analysis will be carried by plotting histogram and scatter plot for the error terms to validate the assumptions

- Residues distribution is normal distribution with zero mean value.

- Residues or error terms of variables are independent

5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes? (2 marks)

Ans. 1. temp coefficient is high and +ve shows Registrations are in high temperatures due to clear weather.

2. light coefficient is 2nd high and -ve shows Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds weather is not comfortable for bike riding hence bike registrations are impacted.

3. yr coefficient is 3rd high and +ve shows Bike sharing registrations are increased from 2018 to 2019 in covid season as people are preferred to use bikes than public transport.

4. Wind speed coefficient is 4th highest and -ve shows bike riding is difficult hence registrations are impacted.

5. spring coefficient -ve shows In USA During the spring, temperatures begin to warm up and thunderstorms and rainstorms are common across the country hence bike registrations are negatively correlated.

6. Summer coefficient +ve shows In summer more people prefer to ride on bikes.

7. September coefficient +ve shows In September month bike registrations are more because more pepole prefer vacation trips on bikes.

8. Mist coefficient -ve shows mist weather is not comfortable for bike riding hence bike registrations are impacted.

General Subjective Questions

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

Ans. linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the Sum of Squared Residuals Method

Assumptions of Linear regression Model

1. Independent variables (input variables) have linear relationship with dependent or output variable.
2. Residue terms or error terms are normally distributed with zero Mean.
3. Residue terms or error terms are will have same variance and independent to each other implies the data homogeneity or homoscedasticity.
4. The independent variables are measured without error.
5. The independent variables are linearly independent of each other, i.e. there is no multicollinearity in the data.

6 Equation: $Y=\beta 0+\beta 1X1+\beta 2X2+...+\beta pXp+\epsilon$

Y= Dependent or output variable, ϵ = error

$\beta 0$ = Y intercept or constant,

X1, X2...Xp = Independent variables.

$\beta 1,...\beta p$ = Co-efficient of independent variables.

Coefficients are obtained by minimizing the sum of squared errors, the least squares criteria

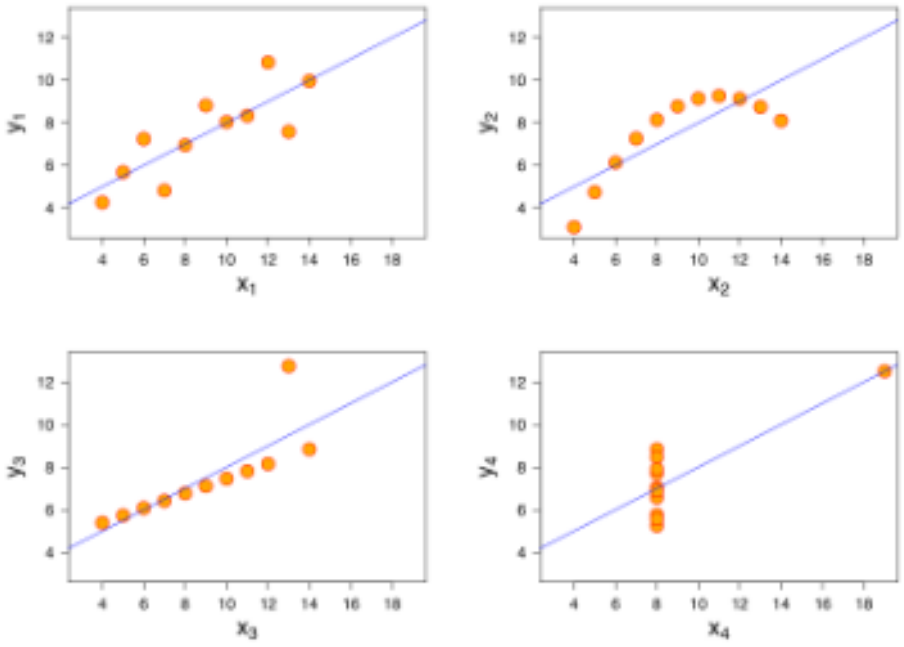
RSS = e1^2 + e2^2+……en^2

e = Y - Ypred

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

Ans. The quartet 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. The datasets are as follows.

These **phenomena** can be best explained by the Anscombe's Quartet, shown below:



As we can see, all the four linear regression are exactly the same. But there are some peculiarities in the datasets which have fooled the regression line. While the first one seems to be doing a decent job, the second one clearly shows that linear regression can only model linear relationships and is incapable of handling any other kind of data. The third and fourth images showcase the linear regression model's sensitivity to outliers. Had the outlier not been present, we could have gotten a great line fitted through the data points. So we should never ever run a regression without having a good look at our data.

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

Ans. Pearson's R correlation coefficient which is the correlation coefficient in the linear regression model. This correlation coefficient is designed for linear relationships.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling

and standardized scaling? (3 marks)

Ans. Scaling is the factor which is used to represent the object size. The size of the object can be shown by increasing or decreasing its original size.

Having features with varying degrees of magnitude and range will cause different step sizes for each feature. Therefore, to ensure that gradient descent converges more smoothly and quickly, we need to scale our features so that they share a similar scale.

StandardScaler is used to resize the distribution of values so that the mean of the observed values is 0 and the standard deviation is 1.

Normalized scale or Min-Max scale shrinks the data within the given range, usually of 0 to 1.

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

(3 marks)

Ans. VIF (Variance Inflation Factor) = 1 / (1 – Ri^2)

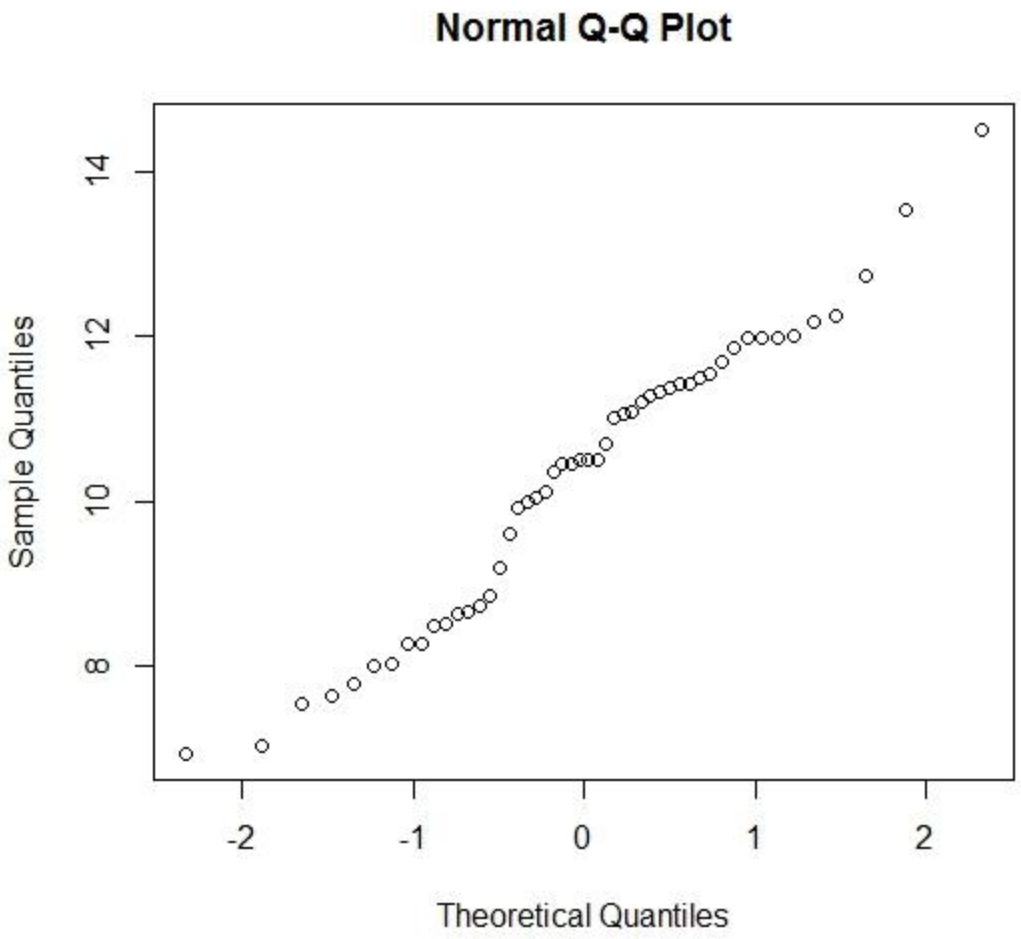
Ri is correlation between independent variable. VIF will become infinity when RI squared value is one i.e both independent variables have perfect correlation. If VIF is 5 corresponding Ri squared is 0.8 which is very high correlation leads to Multicollinearity effect on model hence If VIF>5 corresponding variable should be dropped.

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

(3 marks).

Ans. Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

In Linear Regression Models QQ plot (or quantile-quantile plot) draws the correlation between residues or error terms of independent variable distribution and the normal distribution. A 45-degree reference line is also plotted. QQ plots are used to visually check the normality of the data ie residues.
The normal distribution is symmetric, so it has no skew (the mean is equal to the median). On a Q-Q plot normally distributed data appears as roughly a straight line (although the ends of the Q-Q plot often start to deviate from the straight line). .



QQ plots take sample data, sort it in ascending order, and then plot them versus quantiles calculated from a theoretical distribution. The number of quantiles is selected to match the size of sample data. While normal QQ plots are the ones most often used in practice due to so many statistical methods assuming normality, QQ plots can actually be created for any distribution.
In R, there are two functions to create QQ plots: qqnorm() and qqplot().
qqnorm() creates a normal QQ plot. You give it a vector of data, and R plots the data in sorted order versus quantiles from a standard normal distribution.

Classic bell-curve standard normal distribution with a mean of 0. The 0.5 quantile, or 50th percentile, is 0. Half the data lie below 0. That's the peak of the hump in the curve. The 0.95 quantile, or 95th percentile, is about 1.64. 95 percent of the data lie below 1.64. The following R code generates the quantiles for a standard normal distribution from 0.01 to 0.99 by increments of 0.01.