Sakti Saurav

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)
Ans. There were 4 categorical variables.

1. Season- Customer preferred riding bikes more in Summers, Fall and Winters
2. Months- Mnths 6-10 are influencing more for bike riding
3. WorkingDay - This variable doesn't influence much as there's a similar and colinear variable as well, i.e Holiday.
4. Weathersit - Customer preferred cleared weather compared to mist and rainy.
5. Weekday - Sun, Sat, Thur, Fri are preferred more than other days.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
Ans. drop_first=True, ensures that it removes the extra column created from the category levels. That is if n levels are there, it ensures the first level is not included and n-1 dummy variable columns are returned. The removed column variable can be inferred from the included n-1 variables that is all n-1 0s infer that removed column is True.

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

1. 'Registered', then 'casual', then 'temp'.

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

1. Saw the residual error distribution, the mean was 0 and it followed a Normal Distribution
2. Checked for Homoscedasticity- no specific patterns, and the residual seem to have a constant variance. Plotted residual vs dependent variable in training set.
3. Checked for Multicolinearity- Correlation graph and VIF
4. Checked for linear relation between independent and dependent variables, using CCPR plots. Component and Component plus residual plot

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

1. Temperature (Temp) Customers prefer Summer and pleasant temperature for bike booking
2. Year (yr)- Increase of sales in 2019 compared to 2018.
3. Season, It seems that customers prefered Summer and Winter season.
4. Weather- Customer avoid during Misty and rainy weather and preferred clear weather

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
Linear regression is a model where the target is dependent on single or multiple variables linearly.
i.e. y = m1x1 + m2x2 + …. C1, where y is dependent on x1, x2 etc. where m1, m2 etc are coefficients which means for 1 unit of x1 when others are kept constant, y increases by m1.
When the dependent variables are linearly related to features, it means it could be positively or negatively related.
Below are some points related to Multiple linear regression (and similarly to linear as well)
1. Model fits a 'hyperplane' instead of a line
2. Coefficients are obtained by minimizing sum of squared error (Least squares criterion)
3. For inference, the assumptions from from Simple Linear Regression still hold
    1. Linear relationship between X and Y
    2. Error terms are normally distributed (not X, Y)
    3. Error terms are independent of each other
    4. Error terms have constant variance (homoscedasticity)

And for optimal feature selection:

1. Manual Feature Elimination
    a. Build model
    b. Drop features that are least helpful in prediction (high p-value)
    c. Drop features that are redundant (using correlations, VIF)
    d. Rebuild model and repeat
2. Automated Approach
    a. Recursive Feature Elimination(RFE)
    b. Forward/Backward/Stepwise Selection based on AIC (not covered)

It's generally recommended that you follow a balanced approach, i.e., use a combination of automated (coarse tuning) + manual (fine tuning) selection in order to get an optimal model.

The parameters to assess a model are:
1. t statistic: Used to determine the p-value and hence, helps in determining whether the coefficient is significant or not
2. F statistic: Used to assess whether the overall model fit is significant or not. Generally, the higher the value of F statistic, the more significant a model turns out to be
3. R-squared: After it has been concluded that the model fit is significant, the R-squared value tells the extent of the fit, i.e. how well the straight line describes the variance in the data. Its value ranges from 0 to 1, with the value 1 being the best fit and the value 0 showcasing the worst.

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

Ans. Anscombe’s quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

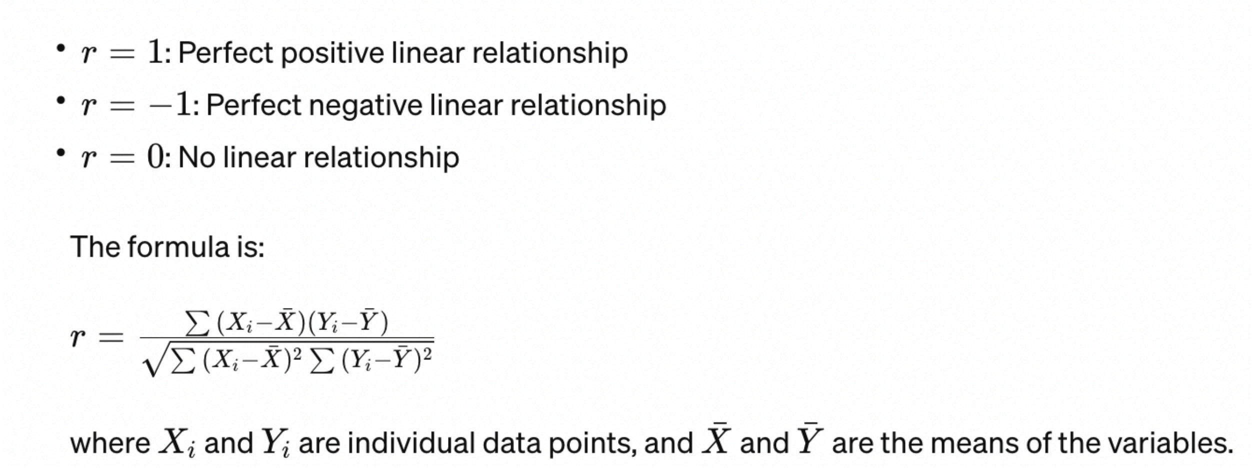
The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

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

3. What is Pearson’s R? (3 marks)

Ans. Pearson’s R is another name of correlation coefficient. Pearson’s Correlation Coefficient, often denoted as *r*, measures the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1, where:



- $r = 1$: Perfect positive linear relationship
- $r = -1$: Perfect negative linear relationship
- $r = 0$: No linear relationship

The formula is:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

where X_i and Y_i are individual data points, and $\bar{X}$ and $\bar{Y}$ are the means of the variables.

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

Ans:

1. Scaling is basically bringing all the variables scale onto same level i.e. range of values of variables would be defined and would be same for al.
2. The purpose is to ensure that all features contribute equally to the model and to avoid the domination of features with larger values. Also, this is required for faster computation
3. Normalisation vs Standardized
   a. Rescales values to a range between 0 and 1
   b. Centers data around the mean and scales to a standard deviation of 1

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
Ans. According to formula of VIF 1 / (1 − R_j^2), where R_j^2 is the score of one independent variable taken as target against other independent variables.
When R_j^2 is 1 i.e. the independent variable is highly correlated with one of the other independent variable, i.e their correlation coefficient is 1. That is when VIF score is infinite.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
Ans. Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

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.

Few advantages:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behavior

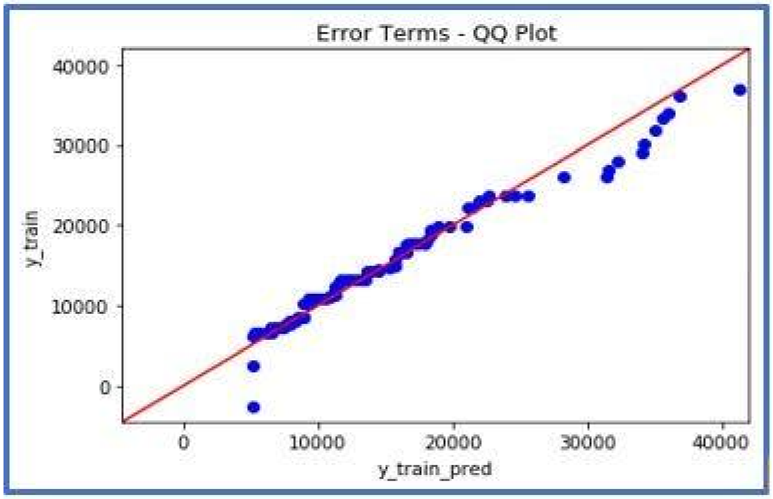
Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

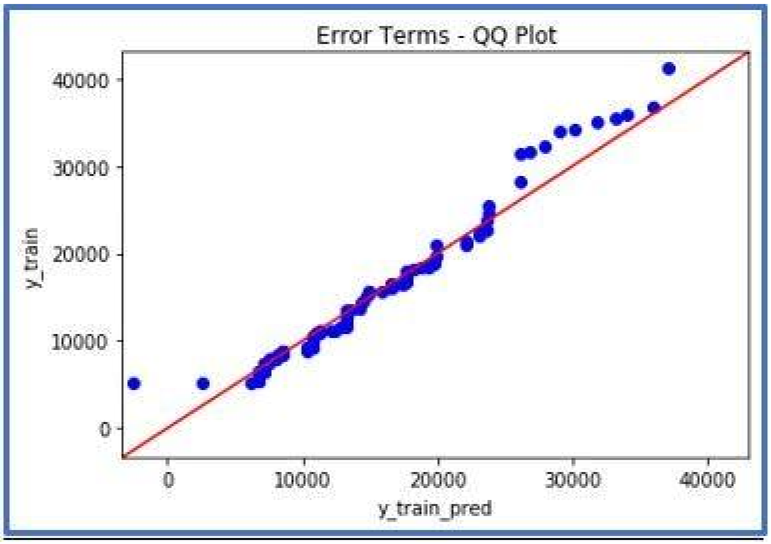
Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis