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?

Inferences gain from categorical variables of the data w.r.t dependant variable

Season: Demand of shared bikes are high in summer and fall season as compared to spring and winter

Year: In 2019, demand of shared bikes is increases as compared to 2018

Month: June, July, August & September are the months where demand of shared bikes is very high.

Weekday: There no strong relation between day of week and bikes demand

Weathersit: Demand of shared bikes are slightly high when weather is clear and less cloudy

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

If we do not use the drop_first = True then we can end up with 'n' number of dummy variables for 'n' number of categories which can cause multicollinearity among these dummy variables also it gives too much same information to model which can lead to redundancy. Due to this redundancy, model will get confused and will not be able to give correct coefficient for each dummy variable. To avoid this multicollinearity and redundancy issue also to improve interpretability of model it is important to use drop_first=True.

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

The feature 'atemp' has the highest correlation (i.e 0.65) with the target variable

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

1. Checked linearity assumption while EDA that independent variable has linear relationship with dependant variable or not
2. Checked Whether distribution of error term is normal with mean 0 or not by plotting distplot of residual (error term = y_act – y_pred)
3. Checked whether there is multicollinearity or not using variance inflation factor (VIF) while modelling

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

'Temp' , 'season_winter' and 'mnth_september' are the top 3 features contributing significantly towards explaining the demand of the shared bikes

General Subjective Questions

1.Explain the linear regression algorithm in detail.

Linear regression is a supervised machine learning algorithm which is used to predict the continuous dependant variable using one or more independent variable. It builds a model which can help us to get a linear equation which shows relationship between target and independent variable.

The main purpose of the linear regression model is to find linear relationship between dependant and independent variable. following is the linear regression equation for one independent variable

$Y = B0 + B1 * X + E$

Y = dependent variable

X = independent variable

B0 = intercept (the value of Y when X is 0).

B1 = slope (how much Y changes for a one-unit change in X).

E = error term (the difference between the predicted and actual values)

Finding the best fitted line which minimizes the error of all the data points is called model fitting.

To find that best fit line will require optimal coefficients of all independent variable which can be calculated by minimizing cost function. Cost function is error between actual and predicted values.

This optimization techniques like gradient descent are used to iteratively update the coefficients of independent variables until the cost function reaches its minimum value. This part is called as model training

After model training, we have to do a residual analysis where we can check the residual (error term) is normally distributed or not for train data.

Once the model is completed, we can make prediction on unseen data by putting the values of independent variables. After predictions we can validate all the assumptions of linear regression to check whether our model is built correctly or not.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a famous example in statistics that illustrates the importance of not relying solely on summary statistics and visualizing data to understand its underlying structure. It consists of four small datasets, which have nearly identical simple descriptive

statistics. However, when you graphically plot these datasets, they reveal distinct and significantly different patterns.

Here are the details of Anscombe's quartet:

Dataset I:

x-values: [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]

y-values: [8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68]

Dataset II:

x-values: [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]

y-values: [9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74]

Dataset III:

x-values: [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]

y-values: [7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73]

Dataset IV:

x-values: [8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8]

y-values: [6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89]

key findings and implications of Anscombe's quartet:

Summary Statistics Can Be Misleading: All four datasets have nearly identical summary statistics, including mean, variance, correlation, and linear regression coefficients. If you only looked at these statistics, you might mistakenly conclude that the datasets are very similar.

Visualization Is Essential: When you plot these datasets, it becomes evident that they have vastly different relationships between the x and y variables. Dataset I shows a linear relationship, Dataset II shows a non-linear relationship, Dataset III shows an outlier-driven relationship, and Dataset IV has an outlier that significantly affects the linear regression line.

Context Matters: Anscombe's quartet illustrates the importance of considering the context of your data. For example, the same summary statistics might describe completely different phenomena, and visualizations can help uncover these distinctions.

Robustness Testing: It highlights the need for robustness testing in statistical analysis. Relying solely on summary statistics can lead to incorrect conclusions. By visualizing the data, analysts can identify patterns, outliers, and other important characteristics that may not be evident from summary statistics alone.

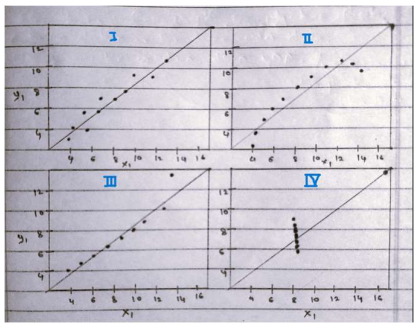

Fig. Anscombe's quartet

In conclusion, Anscombe's quartet serves as a powerful reminder of the limitations of summary statistics and the importance of data visualization in exploring and understanding datasets. It underscores the value of examining data from multiple angles to gain a comprehensive understanding of its underlying structure and patterns.

3. What is Pearson's R?

Pearson's R is a statistical measure of direction and strength between two continuous variables.it is widely use in statistics to assess correlation between two variables.

Pearson's correlation coefficient has 4 characteristics which are as follows:

1.Range: Pearson's r ranges between -1 to 1

An r of -1 indicates that negative correlation means when one variable increases another variable decrease in perfectly linear fashion Whereas an r of 1 indicates that positive correlation means when one variable increases another variable will also increases in linear fashion.

An r of 0 indicates that there is no linear relation between variables

2.Direction: The sign (+,-) indicates the direction of the correlation

Positive r indicates positive relation means increase in one variable cause increase in another variable whereas Negative r indicates negative relation means decrease in one variable cause decrease in another variable.

3.Strength: An r value closer to -1 or 1 shows strong linear relationship whereas r value close to 0 shows less or no linear relationship.

4.Assumptions: Pearson's r assumes that relation between variable is linear also non-linear relationship will not capture correctly by Pearson's r

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

Scaling is a process of transforming data into specific range or distribution to make our modelling process easy. Scaling is performed to ensure that variable with different units should not dominate the modelling process and also it helps us improve stability and performance of various analytical techniques.

Difference between normalize scaling and standardize scaling is as follows:

1.Ranges:

Normalize scaling rescales data between range of 0 to 1 Whereas standardize scaling transforms data to have mean 0 and standard deviation equals to 1.

2.Preservation of relationships:

Normalize scaling preserves relationship between data points whereas standardization does not necessarily preserve the relative relationship between data points

3.Use case:

Normalization is preferred when we want to maintain original scale of the data whereas standardization is use when we want to make data suitable for algorithm that assumes normally distributed data.

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

Variance inflation factor is a measure to assess the multicollinearity in multiple regression model. Multicollinearity occurs when two or more independent variables are highly correlated with each other.

Infinite VIF values occurs when there is perfect multicollinearity present in independent variable which leads to high R-squared value. When value of R-squared is close to 1 then VIF values goes to higher or infinity as calculation of VIF contains 1 divided by (1-R squared).

When R-squared values is close to 1 then denominator value becomes too small which causes very high or infinite VIF value.

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

A Q-Q (Quantile-Quantile) plot in statistics use to measure whether the dataset follows a normal distribution or not. A Q-Q plot compares the quantiles of the observed data with the quantiles of the theoretical distribution.

In Q-Q plot if datapoints are closely follows a straight line suggests that your dataset follows a normal distribution whereas if datapoints are deviate from straight line which means that your dataset does not follows normal distribution.

Importance of Q-Q Plot:

1.Assumption checking:

In linear regression, there are several assumptions one of that is residuals follows a normal distribution. A Q-Q plot helps us to assess that whether the residuals of our regression model follow a normal distribution or not. If Q-Q plot shows a straight line then it means normality assumption is met.

2.Outlier detection:

Q-Q plot also help us to identify the outliers present in data. Outliers may impact our regression model coefficients and predictions. Deviation from the straight line in plot indicates the presence of outliers.

3.Model checking:

Q-Q plot indicates that if dataset does not follow a normal distribution, then linear regression model is not best fit for our dataset and in such cases, we have to look for another regression techniques.

4.Decision making:

Understanding the distribution of residual can be crucial in making decision based on regression model we have. Suppose if residuals are not normally distributed then we should be cautious while making prediction using that model as it may give incorrect predictions.