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?

1.Seasonal Analysis: Fall has the highest average rentals, followed closely by summer.

2.Year-wise Rentals: 2019 sees a notable increase with a median rise of approximately 2000 rentals compared to 2018.

3. Monthly Trend: September tops the monthly rental count, with surrounding months showing substantial demand. The trend aligns with seasonal patterns, indicating a correlation between rentals and seasons.

4. Holiday vs. Working Days: Holidays generally result in lower rental counts compared to working days. Holidays exhibit greater variability in rental demand.

5. Weekday Analysis: Overall, no significant difference in rentals across weekdays is observed. Thursdays and Sundays stand out with higher variability in rental counts compared to other weekdays.

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

drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. If we do not drop one of the dummy variables created from a categorical variable then it becomes redundant with dataset as we will have constant variable(intercept) which will create multicollinearity issue. To keep this under control, we lose one column

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

•atemp and temp both have same correlation with target variable of 0.63 which is the highest among all numerical variables.

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

We validated this assumption about residuals by plotting a distplot of residuals and see if residuals are following normal distribution or not.

I have checked the following assumptions a s well;

•Error Terms do not follow any pattern.

•Multicollinearity check using VIF(s).

• Linearity Check.

•Ensured the overfitting by looking the R2 value and Adjusted R2.

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

•Year(yr)

•Temperature(tmp)

•Windspeed

General Subjective Questions

1.Explain the linear regression algorithm in detail?

Linear Regression Algorithm is a machine learning algorithm based on supervised learning. Linear regression is a part of regression analysis. Regression analysis is a technique of predictive modelling that helps you to find out the relationship between Input and the target variable.

The Equation of linear Regression is y= m1x1 + m2x2 + m3x3 + ………. + m(n) x(n) + c

Where y is target variable and x1, x2, x3 …… xn are predictor variables. And we have two unknowns, m, and c, and we need to choose those values of m and c, which provides us with the minimum error.

We need to get the best fit line which is the line that has the minimum error. In linear regression, when the error is calculated using the sum of squared error, this type of regression is known as OLS, i.e., Ordinary Least Squared Error Regression.

Error function is explained by 'e = - y', and error depends on the values of 'm' and 'c'. Our aim is to build an algorithm which can minimize the error. And in order to do so we use cost function of Linear Regression,

Which is: J (mi, c) = (1/2n)Σ(yi – yp)^2 Where yi and yp are expected values and predicted values. Our main aim is to minimize J by changing m and c and it can be done using Gradient Descent Algorithm. Cost function measures the performance of a Machine Learning model for given data.

2. Explain the Anscombe's quartet in detail?

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.

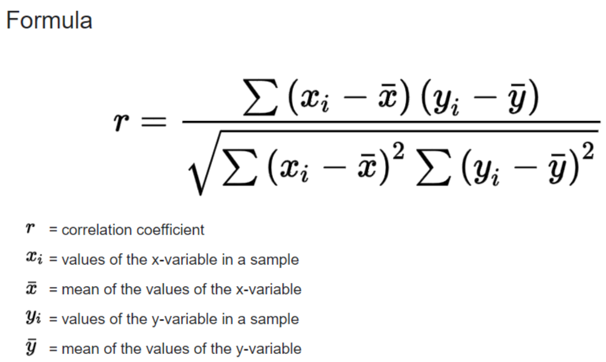
The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

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?

In statistics, the Pearson correlation coefficient (PCC) is a correlation coefficient that measures linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationships or correlations.



Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

When r=1 positve strong correlation

When r=-1 negative strong corelation

When r=0 no correlation

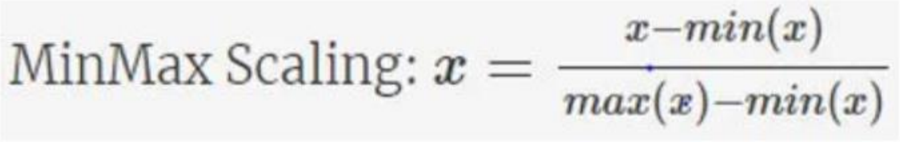
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling

and standardized scaling?

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm. Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. There are two types of scaling

1.Normalization/Min-Max Scaling:

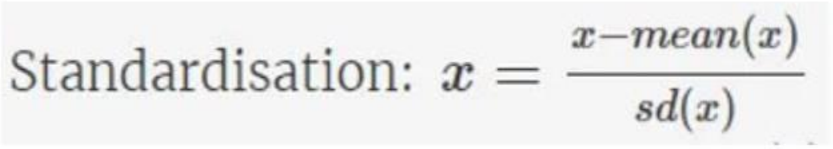
It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.



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

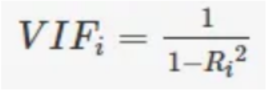
Standardization Scaling:

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).



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

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? The value of VIF is calculated by the below formula:



$$VIF_i = \frac{1}{1 - {R_i}^2}$$

Where, ‘i’ refers to the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

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

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.

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:

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behavior