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?

From visual analysis of categorical variables, following inferences can be made about their effect on the dependant variable

1) Seasons have significant variability in their influence on bike rental demand. Demand is lowest in spring whereas it increases significantly in fall.
2) Median demand for rentals increased by 33% year on year from 2018 to 2019 which may be inferred as a business growth
3) Bike Rentals don't vary much depending on whether a day is a working day or not but here as working days are more compared to non working days it can be inferred that demand per non working day is more.
4) Weather situation as expected has a significant impact on bike rental demand. It is expected that adverse weather conditions should influence and data does exhibit that property. Clear dondtions provide scope for more bike rental demand and snowy conditions significantly reduce demand
5) Bike rentals tend to increase till mid year in both years and then slowly decrease from peak which implies mid year rentals are more

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

When creating a dummy variable, drop_first=True helps in removing one extra column created. This helps in reducing correlation between dummy variables and also conveys the same information. So for example
3 values in a categorical variable can be just identified by 2 dummy variables for example 1 0 and 0 1 represent two values and if it is not both, then it implies 3 values.

Syntax: pd.get_dummies(column_name,drop_first=True)

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

Temp has the highest correlation with count and has a value of 0.63

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

Linear Regression has fundamental assumptions on which model accuracy and reliability depend on. I have validated the assumptions as below

1) Multicollinearity among independant variables should be negligible

Validation: VIF is calculated and any independent variable having a value greater than 5 is removed to maitain no multicollinearity condition.

2)Error terms should be distributed normally
Validation: Distribution of error terms is graphed using seaborn and visual inspection of distribution reveals normality.

3)Homoscedasticity
Validation: Variance of error terms should be constant with respect to independent variables and it is checked by looking at the cone patterns of error terms with respect to x

4)No autocorrellation
Validation: Dependence on the delayed version of itself and is generally present in time series data. No specific test is done by me in this model to check as just two years of data is present.

5) Existence of linear relationship
Validation: Visual linearity should be present among variables and this can be inferred from pairwise scatter plots

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

1)temp variable is highest contributor with a coefficient of .4847 which is significant
2)year is second important contributor with a coefficient of of .238
3)weather_light_snow is next with a coefficient of -.2237(inverse relationship)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent variables, hence it is called linear regression. Since linear regression shows the linear relationship between independent and dependent variables, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables

Mathematically it can be represented as y=a+bx+e

y is a dependent variable

x is independent variable

a is intercept

e is an error term

b is regression coefficient

Types of Linear Regression

Linear regression can be further divided into two types of the algorithm:

Simple Linear Regression

If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

Multiple Linear regression

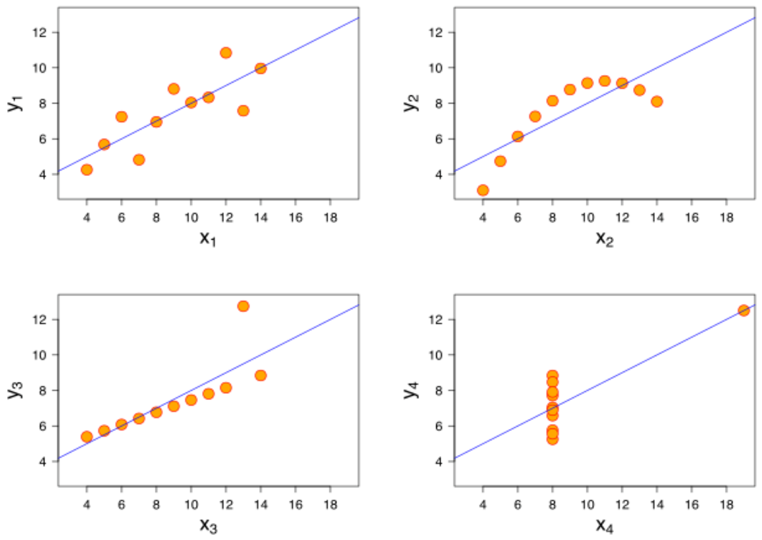
If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Assumptions

1) No multicollinearity between independent variables
2) Homoscedasticity of error terms
3) No auto-correlation in data
4) Relationship between response and feature variables is linear
5) Normality of error terms

2. Explain the Anscombe’s quartet in detail.

Anscombe’s quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points.The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.



- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on *x*.
- The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line. The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

3. What is Pearson’s R?

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases

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

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.

Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

VIF - the variance inflation factor -The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity.(VIF) =1/(1-R_1^2). If there is perfect correlation, then VIF = infinity.Where R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables- If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it’s R-squared value will be equal to 1.So, VIF = 1/(1-1) which gives VIF = 1/0 which results in “infinity”

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

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.It is used to compare the shapes of distributions.A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that’s roughly straight. The q-q plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?