Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about

their effect on the dependent variable?

Answer:

Below are the inferences:

1. Number of Rentals has increased in 2019 compared to 2018.
2. Rental Usage starts increasing from beginning of year and after mid it starts decreasing, Jun to Oct has maximum usages.
3. Usage is comparatively lesser on holidays.
4. Usage is higher when Weather Situation is Clear, in case of Light_Snow its very less while when its heavy rain there is no usage.
5. Mean Rentals are similar on all days, however 25 percentiles, it is lowest on Saturday and Wednesday

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

Answer:

When we create dummy variables for a categorical variable, it changes the values of row to columns and represent in binary form as 1 or 0, 1 being true and 0 being false.

drop_first=true drops one column because data can be still analyzed and same inferences can be made without that column. So, to remove an extra column while creating dummy variables for categorical variable. Also, it reduces the correlation among dummy variables.

For Example:

A column contains Review, it has 3 possible values, “Will Recommend”, “Not Recommend” , “Can't say”.

when we create dummy variable, it will create 3 columns,

but it can be presented using 2 columns as well.

Assume we decide to keep Will Recommend, and Not Recommend

Then Will Recommend Can be displayed as 10

Not Recommend as 01

and Can't Say as 00

So, we drop 3rd variable

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

Answer:

temp (temperature) is having highest correlation with cnt (target) variable.

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

Answer:

We calculated the VIF on the final model and the VIFs were very low indicating less or insignificant multicollinearity.

We also created displot on residuals which indicated errors are normally distributed.

We also created scatterplot to see pattern in residuals and we don't see any pattern so homoscedacity is preserved.

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

Answer:

Temperature, Light_Snow (Weather Sit) and Year

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

Modelling uses machine learning algorithms, in this machine learns from data the way human learns.

Linear regression is one of the most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc. Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression.

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

Types of Linear Regression

Linear regression can be divided into two types:

Simple Linear Regression:

If a single independent variable is used to predict the value of a numerical dependent variable, then it's called Linear Regression.

Multiple Linear regression:

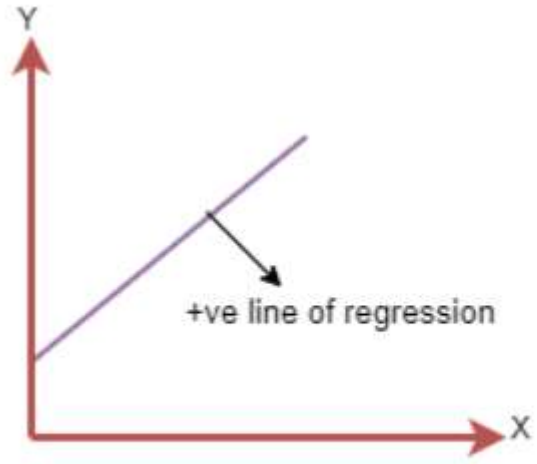
If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Linear Regression Line

A linear line showing the relationship between the dependent and independent variables is called a regression line. A regression line can show two types of relationship:

Positive Linear Relationship:

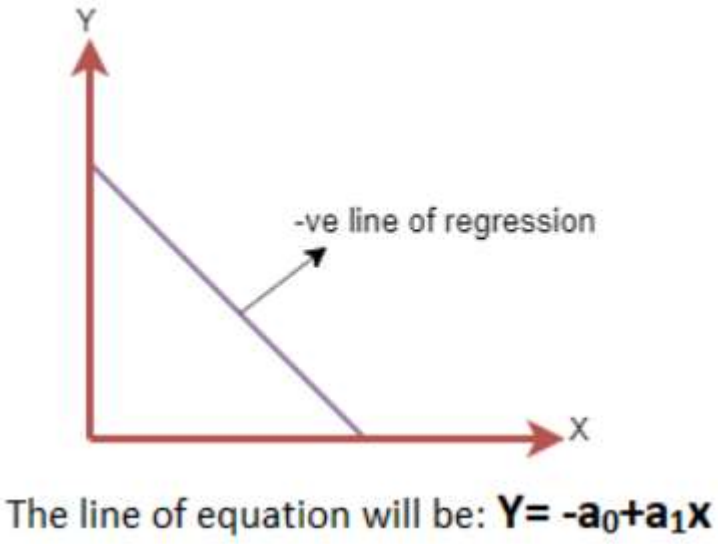
When X increases corresponding to Y, then its Positive Linear Relationship.



The line equation will be: $Y= a_0+a_1x$

Negative Linear Relationship:

When X decreases corresponding to Y, then its Negative Linear Relationship.



Finding the best fit line:

When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least residuals.

Residuals: The distance between the actual value and predicted values is called residual

The different values for weights or the coefficient of lines (a0, a1) gives a different line of regression, so we need to calculate the best values for a0 and a1 to find the best fit line, so to calculate this we use cost function.

Cost function-

The different values for weights or coefficient of lines (a0, a1) gives the different line of regression, and the cost function is used to estimate the values of the coefficient for the best fit line.

Cost function optimizes the regression coefficients or weights. It measures how a linear regression model is performing.

We can use the cost function to find the accuracy of the mapping function, which maps the input variable to the output variable. This mapping function is also known as Hypothesis function.

For Linear Regression, we use the Mean Squared Error (MSE) cost function, which is the average of squared error occurred between the predicted values and actual values. It can be written as:

Gradient Descent:

Gradient descent is used to minimize the MSE by calculating the gradient of the cost function.

A regression model uses gradient descent to update the coefficients of the line by reducing the cost function.

It is done by a random selection of values of coefficient and then iteratively update the values to reach the minimum cost function.

Model Performance:

The Goodness of fit determines how the line of regression fits the set of observations. The process of finding the best model out of various models is called optimization. It can be achieved by below method:

1. R-squared method:

R-squared is a statistical method that determines the goodness of fit.

It measures the strength of the relationship between the dependent and independent variables on a scale of 0-100%.

The high value of R-square determines the less difference between the predicted values and actual values and hence represents a good model.

It is also called a coefficient of determination, or coefficient of multiple determination for multiple regression.

Assumptions of Linear Regression

Linear relationship between the features and target:

Linear regression assumes the linear relationship between the dependent and independent variables.

Small or no multicollinearity between the features:

Multicollinearity means high correlation between the independent variables. Due to multicollinearity, it may be difficult to find the true relationship between the predictors and target variables. So, the model assumes either little or no multicollinearity between the features or independent variables.

Homoscedasticity Assumption:

Homoscedasticity is a situation when the error term is the same for all the values of independent variables. With homoscedasticity, there should be no clear pattern distribution of data in the scatter plot.

Normal distribution of error terms:

Linear regression assumes that the error term should follow the normal distribution pattern. If error terms are not normally distributed, then confidence intervals will become either too wide or too narrow, which may cause difficulties in finding coefficients.

It can be checked using the q-q plot. If the plot shows a straight line without any deviation, which means the error is normally distributed.

No autocorrelations:

The linear regression model assumes no autocorrelation in error terms. If there will be any correlation in the error term, then it will drastically reduce the accuracy of the model. Autocorrelation usually occurs if there is a dependency between residual errors.

2. Explain the Anscombe's quartet in detail.

Answer:

Francis Anscombe illustrated the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.

There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

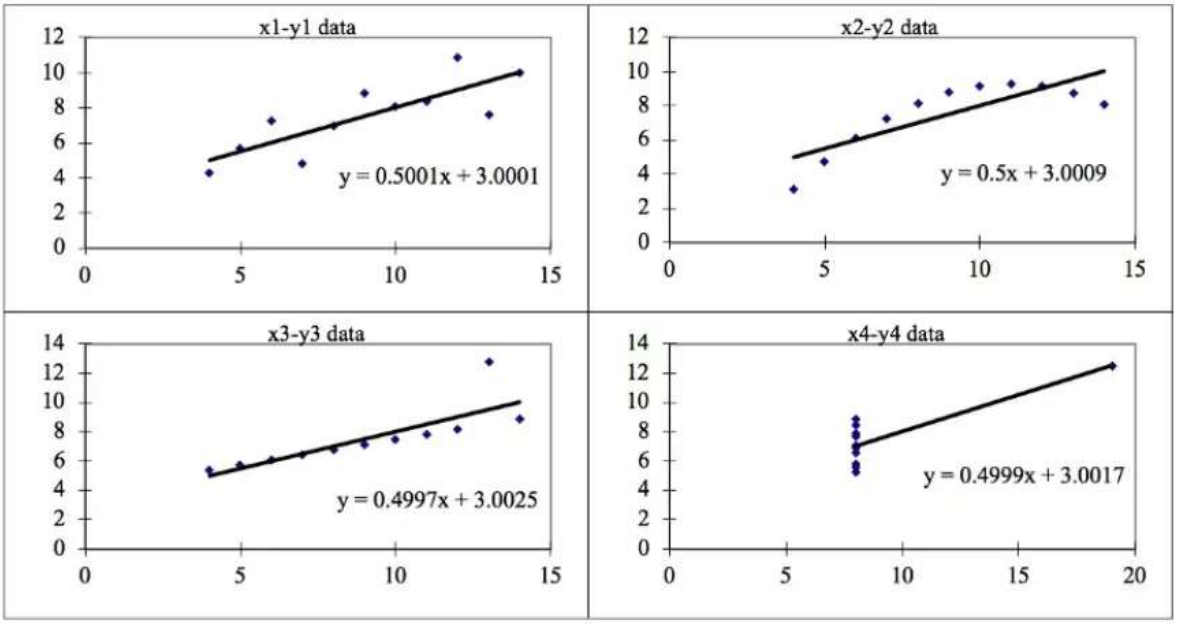
As per him, it's important to visualize data before applying various algorithms to build models. Features must be plotted in order to see the distribution of the samples to identify anomalies present in the data like outliers, linear separability of the data, etc. These four plots can be defined as follows:

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

The statistical information for these datasets is similar and can be computed as follows:

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

When these models are plotted on a scatter plot, all datasets generate a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



The four datasets can be described as:

Dataset 1: this fits the linear regression model well.

Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.

Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model

Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression **model**

So, its important features in the dataset must be visualized before implementing any machine learning algorithm on them which will help to make a good fit model.

3. What is Pearson's R?

Answer:

The Pearson R is the most common way of measuring a linear correlation. It measures the strength and direction of the relationship between two variables and is a number between –1 and 1.

When Pearson R is between 0 and 1, then it represents Positive correlation

When Pearson R is 0, then it represents No correlation

When Pearson R is between 0 and -1, then it represents Negative correlation

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

and standardized scaling?

Answer:

Scaling of data is a preprocessing step which we apply on independent variables to normalize them to fit in a range.

We perform scaling because, data can be in different units and can have very high/low values/range. Since while regression modelling, we use values and not the units of the values and it can give wrong results. So, we do the scaling so all features can have same level of values.

For Example, Age could be between 0 -100, whereas Year can be between 1900 and 2100, so even though they may not be adding so much importance to model, mathematically based on value it will say so.

Scaling does not impact p-values or R-squared values, it just affects the coefficients.

Normalization:

With normalization all values of a column is fit between 0 and 1.

We can use MinMaxScaler for the same.

Formula:

normalization = (x -min_x)/(max_x -min_x)

Standardization

With Standardization all values of a column are replaces using Z-score. It brings all data into standard normal distribution.

Only problem is we lose some info in data especially about outliers

Standardization = (x -mu (mean))/sigma (standard deviation)

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

Answer:

When VIF is infinite it means correlation is perfect, which means the variable can be linearly explained by some other variable. So, it can be removed from model.

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

Answer:

Q-Q plot short of Quantile-Quantile is plot of quantiles of two distribution with respect to each other. Whenever we are interpreting Q-Q plot, we focus on **y=x** line.

The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against.

In linear Regression they are used to check if the points lie on the line. If they are not, then errors may not be normally distributed.

For Example: if we test distribution of age of Employees in team, then it means we are testing quantiles of team members age vs quantile from normally distributed curve.

If two quantiles are sampled from same distribution, then they shall fall in same line.