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)

Solution

From the analysis, following are the categorical variables having impact on dependent or target variable **cnt** (y):

Column	Impact on bike rentals
yr	There is an increase in bike rentals from 2018 to 2019
season	Spring has the least number of rentals recorded
weathersit	Slight changes in weather are impacting the rentals, this can be inferred from box plot (compared to weather conditions 2 & 3, weather condition has high rentals) **Negatively correlated**
mnth	Month September has high rentals

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

Solution

During encoding of any categorical, we tend to use this parameter in pd.get_dummies function or it's alternative in sklearn onehotencoder **drop='first'** to avoid high correlation. For example, if variable has n levels, we can drop any kth level out of n levels which means we don't have to specifically mean or imply the impact of that kth level. Thereby it avoid dummy variable trap as well.

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

Solution

tmp has the highest correlation with target **cnt** variable

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

Solution

Following are the ways to validate the assumptions of Linear Regression after building the model:

- Insignificant variables should be removed and must be validated with summary
- There should not be any variable with high VIF (VIF in simple terms means can be explainable/ correlated with other features)

- Residuals should be normally distributed (with constant variance)

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

Solution

a. Weather situation
b. Temperature
c. Wind speed

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Solution

Linear regression algorithm is a method of finding linear relationships between dependent (target variable) vs independent (feature variables).
Linear regression has 2 subtypes:

- Simple Linear Regression – single dependent and single independent variable is present and it takes a form of “y=mx+c” equation mathematically.
- Multiple Linear Regression – we will have more than one independent variables/ features/ predictors to find a relationship between features and target (equation of type “y=b0x0 + b1x1+b2x2+….+bnxn+c)

A regression algorithms predicts the value by minimizing the squared error distance between actual value and predicted (also called as residuals).
Following are the few such metrics that help in minimizing the error rate for the model:

1. RSS: Residuals Squared Sum
2. Adjusted R-Square: specifically useful when we want to assess the impact of a variable on the model

Following are the steps in building a model, specifically Multiple linear regression:

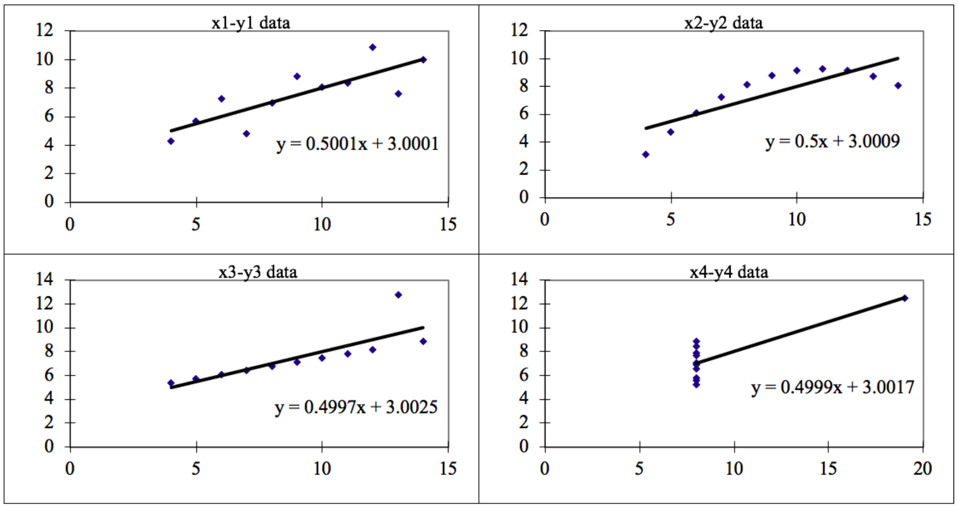
1. Removing highly collinear/ correlated variables
2. Scaling/ Standardization of features
3. Encoding of categorical variables
4. Feature selection either through

a. Significance + VIF (Manual)
b. RFE – recursive feature selection automated
c. RFE + Manual

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

Solution

Anscombe's quartet has four datasets that are nearly identical in descriptive statistics as show in the image. Following image depicts that they have almost similar equation but their distribution is different when plotted as scatter plots



- First plot has linear relationship with data points normally distributed
- Second plot has non-linear relationship and it is not normally distributed
- Third plot has linear relationship has line that fits almost better but because of outlier it has similar descriptive statistics compared with others
- Fourth plot explains that a high value single data point is enough to mean/ represent a high correlation between the variables

3. What is Pearson's R? (3 marks)

Solution

It is a linear coefficient that returns value ranging between –1 and +1. It represents a strength of relationship between the variables.

- -1 means highly and negatively correlation, say example increase in one value can have negative impact on other and have decrease in value
- +1 is vice-versa to the above, increase in one variable can mean increase in other
- A neutral of value 0 means, those variables neither positively or negatively impact each other

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

Solution

Distance based algorithms like Linear Regression are sensitive to numerical variable value ranges. Hence it is recommended to perform scaling so that values will have their essence/ impact on target remain same.

Two types of scaling:

1. Normalized Scaling – brings the value scale between 0 and 1. In other words, values are shifted and rescaled such that any value seen so far will always be in the range of 0 and 1. (formula = (xi – min_x)/max_x-min_x)
2. Standardized Scaling – values are centred around mean and with a unit variance/ standard deviation

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

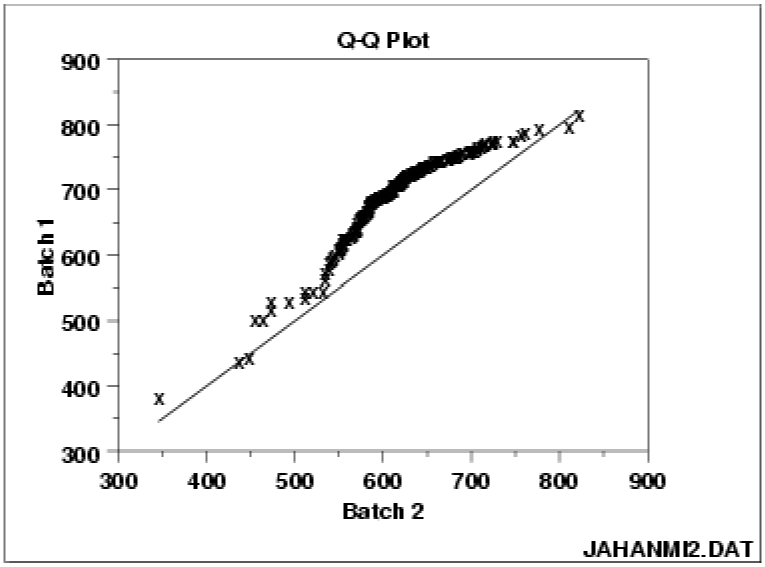
Solution

VIF value is infinite when values are highly correlated, it is calculated as inverse of R-square subtracted from 1. VIF can only be infinite when it’s R-Square is 1 which means other features will be able to explain the variance in a specific variable

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Solution

A Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.


Q-Q Plot
Batch 1
Batch 2
JAHANMI2.DAT