3/19/2019

Shubham handa

8638369

ASSIGNMENT 3

MATHEMATICS FOR DATA ANALYSIS

Q1)

1. In order , to summarize the three variables in the question the nationality was divided into two categories. This makes easier for us to divide Canadian amongst non-Canadian nationalities. Then using summary and only selecting the variables asked we could calculate and differentiate between the two nationalities.
2. In order To summarize the three variables the question has asked the countries were divided into Canadian countries were given 1 priority number and usa country teams were given 1 number. Then using these conditions, the data was summarized by summarize and summary. Also note, there were two players who played for both usa and Canada as they played for Canada first, they were assigned to Canadian countries.

And calculated according to that.

1. To solve this problem a lot of methods could be used at first the differences between most of the factors calculated could give us a brief idea. To go more into detail, a lot of comparison tests could be used. for this purpose, at first the normalcy of each variable was evaluated. If the variable was normally distributed t test was run and if the data was not normally distributed Wilcox test was used (in the case of GAA). Also graphical methods could be used such as plotting graphs and evaluating if they are close and different to each other. for this boxplots and plots were used. The reason for choosing this was as these graphs can divide into categories and help us clearly identify if the data is related to each other or not. Also , box plot can help us identify outliers , but that comes in the later part of the dataset. After this to get more of an idea linear regression was used to evaluate the relation of one to another. As you can see from the graphs in some cases the data is not that significant (coefficient and intercept). Also in tests If the p value was >0.05 then the data was not that different. Otherwise there could be a significant difference for data. The P values are calculated in the R script.

Q2)

1. The next part of the question was to transport the data and only use variables which would help us further analyze the dataset. There was a condition in the requirement tough, only those games were selected where the games played were greater than 2.
2. Also, all the data was converted into numeric values. The reason for this was regression analysis only works on numeric values. So, all the data was converted into numeric values (variables such as salary).

3) As said earlier the requirement was to take goalies who had played more than 2 games.

All the coloums were transferred into a new table and this table was used in the subsequent analysis.

Q3)

1. All the suitable univariant graphs were calculated using multiple graph equation i.e if the variable is numeric a histogram will be resulted.
2. To calculate correlations multiple correlation method was used to calculate the correlation for all the variables also ,same was calculated using cpairs and the library used was gclus for this. This allows to yield a graphical result of multiple correlations compared to each other in colored format.
3. Before calculating the outliers, each variable was tested for normalcy using shapiro-wilko test and the result was plotted. The reason for doing this was to get an idea how closely values lie to each other and if there is some value which is making a huge difference. (QQ plot) Then the outliers were evaluated by plotting each variable in a box plot. This can easily help us find the no. of values that are outliers and then we can find the values that are outliers. I calculated the outliers using traditional method of using IQR and calculating them then I put this into a table listing all of them stating if the values are outliers or not. (yes or no)
4. The thing about treating outliers varies data to data i.e you need to closely evaluate data sometimes the data might be correct (for eg: Canadian player getting huge salary) in that case you cannot just remove or treat the data. Also, for some missing values in 2 cases you need to either remove them or replace them with null values. In this case I went with the later and replaced them with null values. The reason for doing this was it was not making that significant difference.

Q4)

1. The 4 type of variable selection methods used for salary were

Forward selection, backward selection, stepwise criteria selection and baseline model selection.

1. The reason for doing variable selection or feature selection is that they help us in determining which variables have the best affect on the data and which have the least. This can be seen using graphically and numerically plotting them. Then I created models and residual vectors and plotted them.
2. Shapiro wilko’s test helps us determine how strong our model is The more significant the p value the better and the less error that is the more data is correct and our justification is correct. As you can see from the r script the p value is >0.05 , Hence, our data is strong. Also, we can say our data is strong if it is independent of each other , meaning if the variables are independent of each other. Chi square test can help us identify that and the results were shown out. The strong variables that came out were , dob , nationality , height and weight.

Q5)

1. The 4 type of variable selection methods used for salary were

Forward selection, backward selection, stepwise criteria selection and baseline model selection.

1. The reason for doing variable selection or feature selection is that they help us in determining which variables have the best effect on the data and which have the least. This can be seen using graphically and numerically plotting them. Then I created models and residual vectors and plotted them.
2. Shapiro wilko’s test helps us determine how strong our model is the more significant the p value the better and the less error that is the more data is correct and our justification is correct. As you can see from the r script the p value is >0.05, Hence, our data is strong. Also, we can say our data is strong if it is independent of each other, meaning if the variables are independent of each other. Chi square test can help us identify that and the results were shown out. The strong variables in this case that came out were height weight , salary , shoots.