Name:- Swapna Vaidya ***Class:-*** MSDS - 422

EXPLORING AND VISUALIZING DATA

(1) ***A summary and problem definition for management***

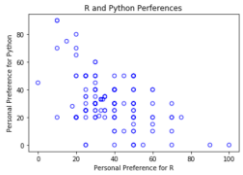
The Northwestern University MSPA department in year 2016 conducted a survey to help them understand if they should be updating the program with new classes and the software that they currently use is in demand or not. The survey got 207 respondents. This data was given for exploring to help management with the following objectives: -

- Learn about current student software preferences.
- Learn about student interest in potential new courses.
- Guide software and systems planning for current and future courses.
- Guide data science curriculum planning.

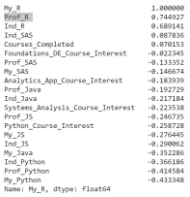
(2) ***Discussion of the research design, measurement and statistical methods, traditional and machine learning methods employed***

Based on the exploratory data analysis that was done the survey data, below were some of the findings:-

- In all there are 207 instances with 40 features in the data
- Of the 40 features, 20 features are non-numerical
- There are missing values in the columns related to the 4 new courses and the number of courses taken. These instances would need to be taken care of before doing any modeling.
- As per the histogram, the data for software like Java, JavaScript and SAS from industrial, personal preference and Professional are right skewed with a long tail. This indicates that there is not much interest in these courses from a personal, industrial and professional perspective for the students.
  While for software languages as Python and R from industrial, personal preference and professional is normally distributed. This indicates that there is lot of interest in these languages from a personal, industrial and professional perspective for the students.
  Similarly, there is a lot of interest in the 4 new courses. The Python course is left skewed with a long tail, with the maximum interest in it.
- As per scatterplots for personal preference between R and Python, there seems to be a negative correlation between the 2 software's. Students who prefer Python the most, don't have much interest in R. Meanwhile students whose interest in R increases their interest in Python decreases. While there are quite a few students who are interested in both the languages.

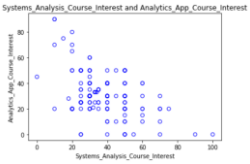


- Based on the correlation matrix of the feature "Personal Preference Of R" with other features, there is high positive correction between Industrial Preference of R and Professional Preference of R. At the same time a high negative correlation between personal, Industrial and Professional preference of Python.

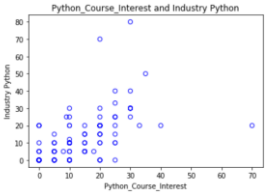


```
My_R                                1.000000
Prof_R                              0.744927
Ind_R                               0.689141
Ind_SAS                             0.087836
Courses_Completed                   0.070153
Foundations_DE_Course_Interest     -0.022345
Prof_SAS                           -0.133352
My_SAS                             -0.146674
Analytics_App_Course_Interest      -0.183939
Prof_Java                          -0.192729
Ind_Java                           -0.217184
Systems_Analysis_Course_Interest   -0.223538
Prof_JS                            -0.246735
Python_Course_Interest             -0.258728
My_JS                              -0.276445
Ind_JS                             -0.290062
My_Java                            -0.352286
Ind_Python                         -0.366186
Prof_Python                        -0.414584
My_Python                          -0.433348
Name: My_R, dtype: float64
```

- For the 4 new classes, there seems to be a negative correlation between System Analysis Course and Analytics App Course. Students very interested in System Analysis Course are not interested in taking the Analytics App course.



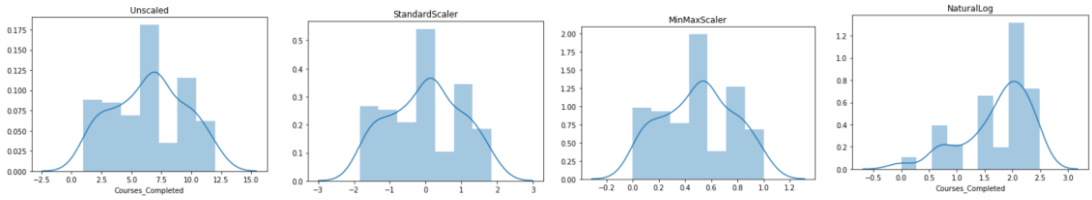
- Another interesting insight that was gained as part of the scatter plot between the Python Course and the Software was that there is a positive correlation between Python Course and the Industry Python interest. This is very true and inline with the current industry demands.



(3) ***Overview of programming work***

This assignment was done in Google Colab, using 4 main libraries (Pandas, numpy, matpotlib, Seaborn and Skylearn).

- ***Ingestion:-*** Using pandas the data was first loaded from Google OnDrive. The "RespondentID" was used as an index for the data. The feature names were renamed to be more friendly.
- ***Exploration:-*** of the data was done using methods like describe(), info() and corr().
- As part of exploration, ***visualization*** was done using histograms and scatterplots.
- ***Data Cleaning and Transformation:-*** Next the data was then split into train and test data using the "train_test_split" method of the Skylearn package. Then using the Skylearn package class "Mutator", the trained data was cleaned with the "median" strategy. Since all the numerical features except the "Courses Completed" are in the scale of 0 to 100, the Standard Scalar and MinMax scalar method was looked into. Without being scaled, the "Courses Completed" seemed normally distributed with no outliers etc.. Either using the Standard Scalar method or the MinMax Scalar method resulted in the data being normally distributed with different scales. Using the LogMethod resulted in a left skewed curve. So to ensure all the features are scaled, the Standard Scalar method was used.

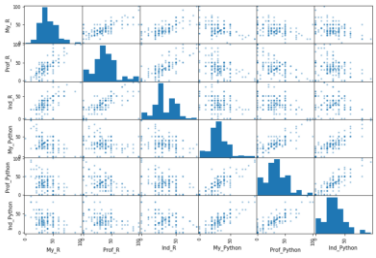


- ***Modeling:-*** The data was then separated into label and features values. The label here was defined for the "new Pyhton Course". Next a simple linear regression model was used to fit the label and feature values. The MPE on the train data was 25.96, MAPE is 20.68 and an R^2 of 0.25. These metrics indicate that model is not very good with high MAPE, an overbias in the accuracy and very low R2.
- ***Evaluation on the Test Data:-*** Next the model was evaluated on the test data to ensure there is no overfitting or underfitting of the model. The Imputer strategy of "Median" and the scalar method was transformed on the test data. It was not fitted but only transformed based on the fit values from the train data. The linear model was then used to predict the labels. Based on the MPE, MAPE and R^2 of the test data the errors were in line with the train data, hence it was ensured that there was no overfitting of the data. Having said that more models need to be explored to fit the survey data.

(4) ***Review of results with recommendations for management***

After the exploratory data analysis, following are the recommendations for the NorthWestern MSPA management department:-

- The management should update the software to R and Python. They should offer the classes in R and Python, as students are interested in either both the languages or learning one or the other language. But the interest in Java, Javascript and SAS is minimal, hence the classes should not be offered in these 3 languages.



- The management should consider offering the 2 courses **System Analysis** Course and **Analytics App** Course as electives as there is a negative correlation between the 2 classes. This will enable the students interested in one or other course to take the courses as they prefer.
- The Management should definitely offer the Python Course not as an elective but as a core course as there is a definite correlation between what the students see the importance/prevalence of Python in their industry and the student's interest in taking the new course if offered by the university.