**Term project:  
Students’ Performance EDA**

Mohammed majjaj

DSC530-T303 Data Exploration and Analysis (2213-1)

**Summary:**

**Outcome of EDA**

**An outlier** in our data we see a consistent few data points where the score is below 25 and away from the bulk of the data around 40 and 90.

they are legitimate observations that accurately describe the few students that scored low in some subject. it will need more analysis to understand the why and if there is any correlation with other variables in this study.

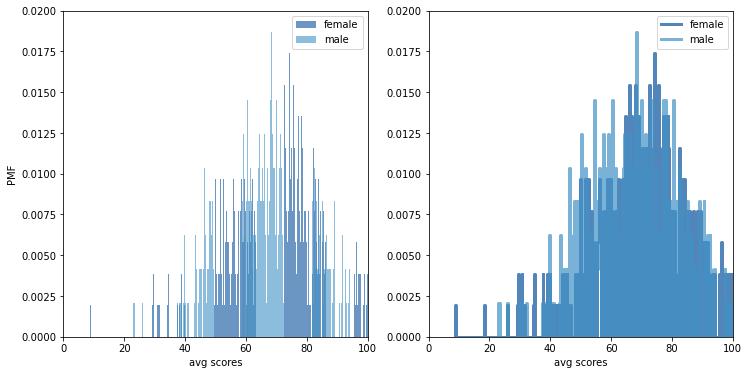
**Mean:**  
we can see that the average student scored in the sixties point, in this order from highest to lowest. reading score, writhing then math score. we can see that the average student scored higher in avg language score then math score. and the overall avg score is 67.77

**Mode:**  
the most frequest score in math is 65  
the most frequest score in reading is 72  
the most frequest score in writing is 74  
the most frequest score in avg Language are 67, 73, and 75.5  
the most frequest score in overall avg score is 68

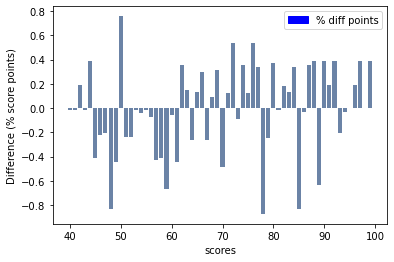
**spread:**  
we can see that the spread is very high the difference between max and min value for math as an example is 100 point. we can see that across all variables the spread is high, and the Standard variation confirm this observation.

**histograms shape and tails:**  
all variable's histograms above have a bell-shaped picture, usually presents a normal distribution. with some skewness to the left where large number of occurrences in the upper value cells (right side) and few in the lower value cells (left side).

**comparing the overall avg score between female and male: calculating the PMF for avg score female and male.**

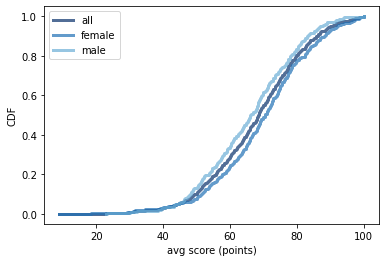
****

we can see that from the chart that female is less likely to score lower than 60. and somewhat more likely to score higher than male.



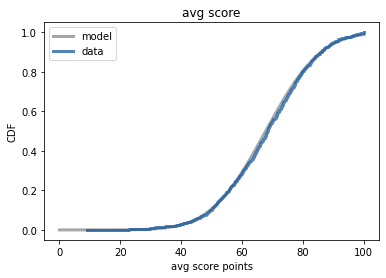
this chart shows the difference in percentage point between the two Pmfs (female\_pmf and male\_pmf). it chows the result as bar chart, it makes the pater clearer: female are less likely to score lower then 60. and somewhat more likely or to score higher than male.

**calculating the CDF for avg score for all class then female and male separately for comparison.**

****

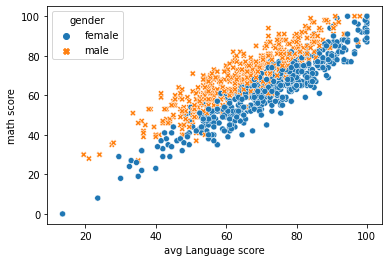
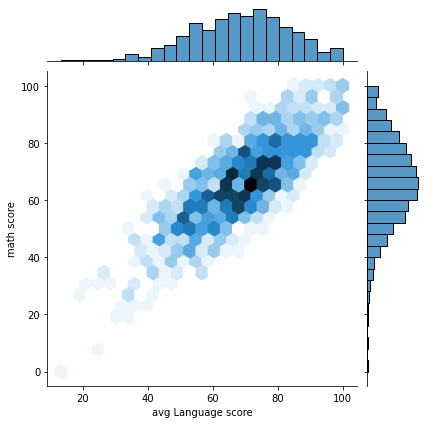
this figure makes the shape of the distributions, and the differences between the different groups apparent. we can see that the scores for male are consistently lower throughout the distributions after 50 average score point.

**compare a model of a normal CDF to our Data average score CDF**

****

we can see that the data fits well the normal model.

**Scatter plot of avg Language score and math scores**

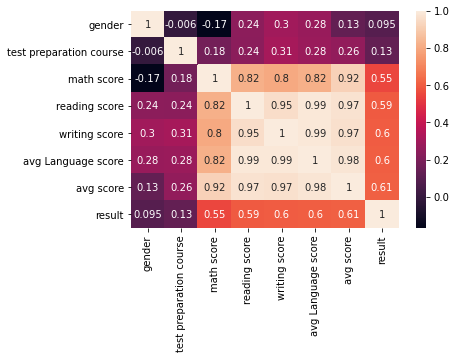
****

we can see that the avg Language score and math score have positive linear relationship; when a student has a high score in Language score it is expected to have a high score in math score.

there is also a distinction of the degree of the relationship. we can see that there is a clear distinction between the two genders in second graph.

for female student for a math score the corresponding language score is lower. for male student when math score is higher the corresponding language score is lower.

**correlation matrix**

****

we can see from this correlation map that:

gender female has a negative point biserial correlation with Math score which we already have noticed from scatter plot above

test preparation course has a positive point biserial correlation with average score and suggest that student that conducted preparing have positive results

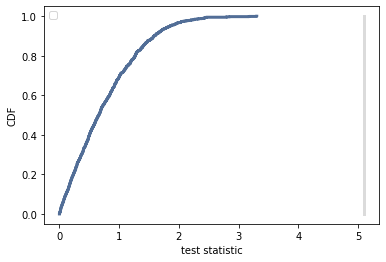
the writing score has strong positive correlation with reading score.

the math score has a strong positive correlation with writing and reading score but has a stronger correlation with avg score.

average score has a positive correlation from strongest to less strong, language score, math score, test preparation, and gender.

**Conducting a hypothesis test**

* test the assumption we made earlier that student gender affect the Math score. will be using as a test statistic the difference in means between both groups.
* our null hypothesis is that the apparent affect is not real (gender do not affect math score).

****

The reported p-value is 0 (statistically significant). which means that in 1000 trials we didn't see a difference in score mean between the two groups, under the null hypothesis, that exceeded the observed difference. That means that the p-value is probably smaller than 1/1000 , but it is not actually 0.

To get a sense of how unexpected the observed value is under the null hypothesis, we can compare the actual difference to the largest value we saw in the simulations.

**What do you feel was missed during the analysis?**

I could have utilized other machine learning algorithm to model and predict the variable gender in addition the logistic regression I utilized in my analysis and compare between the outcome and accuracy.

**Were there any variables you felt could have helped in the analysis?**

I would be interested to see the relationship between the parental level of education, test preparation course, and the score of students. Would be interesting to see if there is a significant difference on average score or it is only a matter of chance.

**Were there any assumptions made you felt were incorrect?**

The assumption I made that female tend to score lower in math score than male. The hypothesis test based on test statistic the difference in means math score between bot groups showed that the apparent difference is statistically significant and that it was not a result of chance.

**What challenges did you face, what did you not fully understand?**

When I tried to apply the logistic regression to predict gender variable using more then 6 variable the code breaks with an error I did not understand and had to reduce the number in formula to field the model so it can work. I looked on the internet for a solution and most of the answer explain the error that it is because Maximum no. of iterations has exceeded the limit. The proposed solution didn’t work for me and had to reduce the explanatory variable I’ve used to only 6.