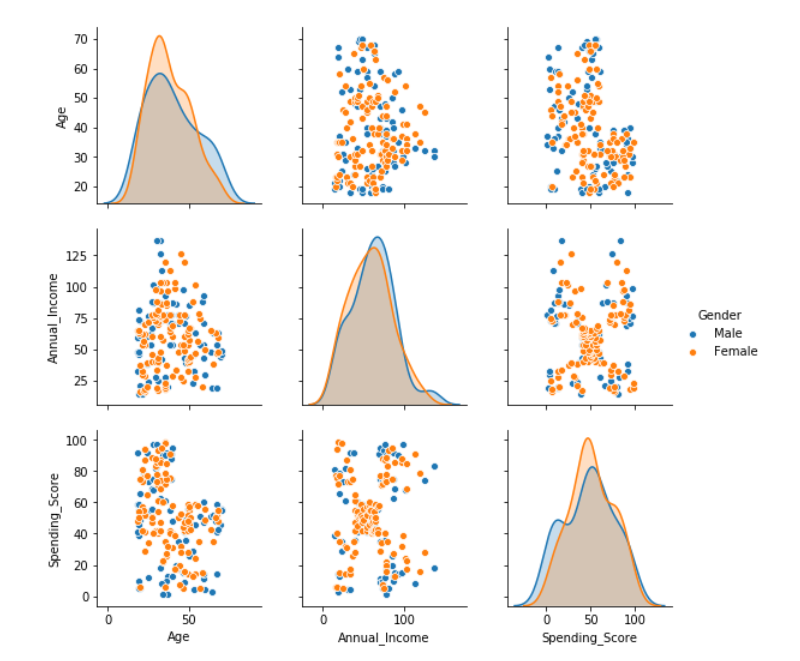
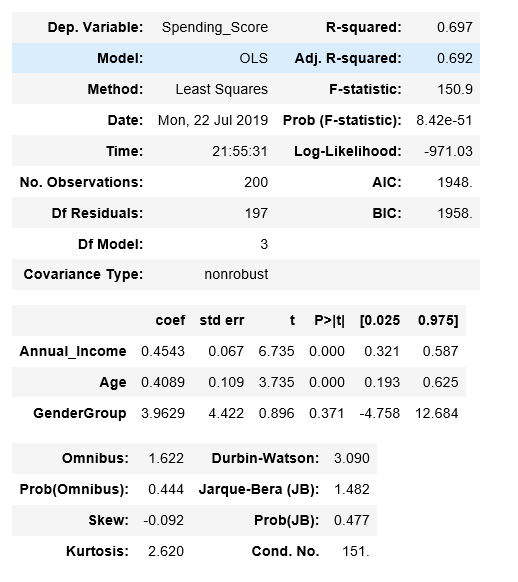
Data source: <https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python>

Business case: see the link above.

Summary of observation on the original data set:





1. No clear linear relationship between the spending \_score and independent variables like annual\_income and age.

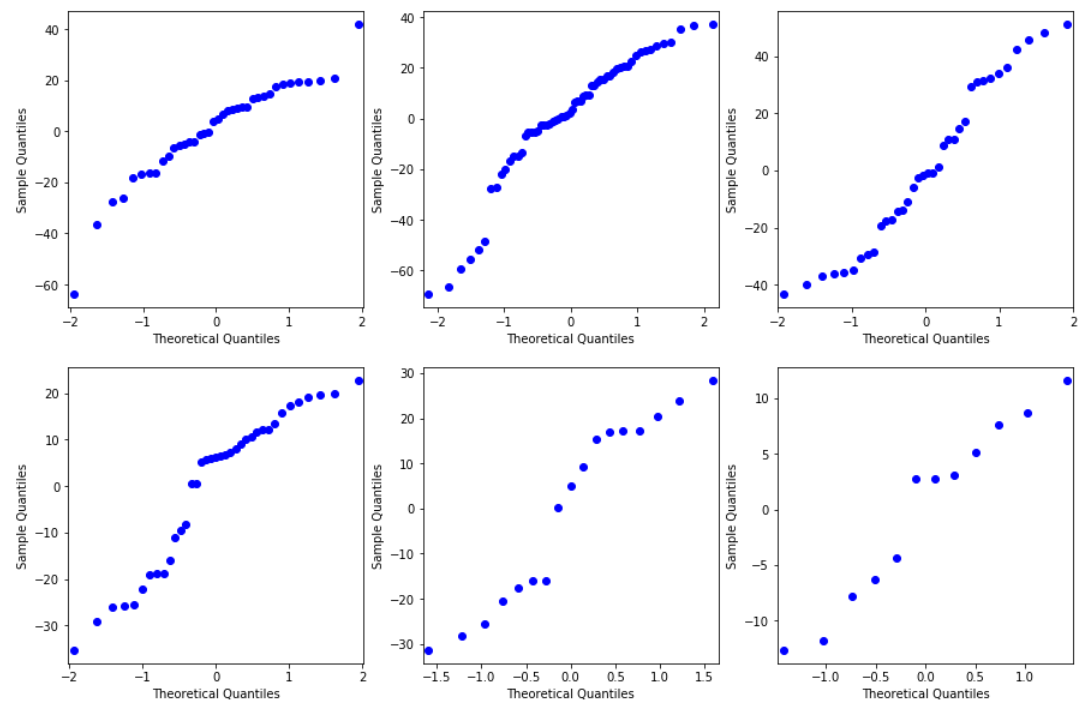
2. A single linear multivariate regression that models the relationship between the dependent variable and the independent variables returns a low R-squared at around 70%

3. An intercept should be excluded as no annual income suggests no spending power/score.

4. …… (feel free to add more stuffs related to data inspection like checking if there is any #NA, mean of different variables etc.)

Model: apply a forward elimination on the linear multivariate regressions on each age group to select the combinations of independent variables that yield highest R-squared.

|  |  |  |
| --- | --- | --- |
| Age Group | Regression | R-squared |
| Below 25 | Spending\_Score = 4.0296 \* Age - 0.6769 \* Annual\_Income  (T-value:0) (T-value:0) | 89% |
| 26 - 35 | Spending\_Score = 1.6762 \* Age + 0.1718 \* Annual\_Income  (T-value:0) (T-value:0.11) | 86% |
| 36 - 45 | Spending\_Score = 0.5828 \* Age + 0.2917 \* Annual\_Income  (T-value:0.13) (T-value:0.15) | 70% |
| 46 - 55 | Spending\_Score = 0.7401 \* Age  (T-value:0) | 83% |
| 56 - 65 | Spending\_Score = 0.5358 \* Age  (T-value:0) | 71% |
| Above 66 | Spending\_Score = 0.2173 \* Age + 0.6344 \* Annual\_Income  (T-value:0.01) (T-value:0.19) | 97% |



Observation:

1. R-squared for regression on each age group improve significantly.

2. the residuals are normally distributed.

3. Some independent variables have T-value higher than 5%, suggesting that they are not a significant determinant for the spending\_score. However, we argue that including those variables give us the best predicting power as combing them with other variables yield highest R-squared.

4. Annual\_Income is negatively correlated with the spending\_score for customers below 25.

5. Annual\_Income does not seem to have any relationship with the spending\_score for customers between 46 years old to 65 year old.

6. there are not many data for the age group above 56 (56-65 and above 66). Therefore, the regression results in these two groups could change significantly with more data collection.