COMP8430 Assignment 1

Longfei Zhao, u5976992

Question 1

1. Yes, I think the problems and issues raised in the paper are still relevant today. Even today in some cases we still need to manually write data, which will cause problems mentioned in the paper.
2. Maybe how to deal with other types of data is a new problem, such as picture, voice. For example, in some system, we need users' picture to do face recognition. How do we make sure the format of those picture is correct is a issue.

Question 2

Split my ANU ID into 59, 76, 99, 2.

Therefore, L = {5, 20, 32, 15, 42, 55, 92, 84, 56, 42, 13, 74, 59, 76, 99, 2}.

Sort L, get {2, 5, 13, 15, 20, 32, 42, 42, 55, 56, 59, 74, 76, 84, 92, 99}.

1. the median of L

   There are 16 numbers. Therefore, the median is the average of the 8th and 9th, which is (42+52) / 2 = 46.
2. the mode of L

   42 appears twice. Other numbers all appear once. Therefore, the mode of L is 42.
3. the five number summary of L

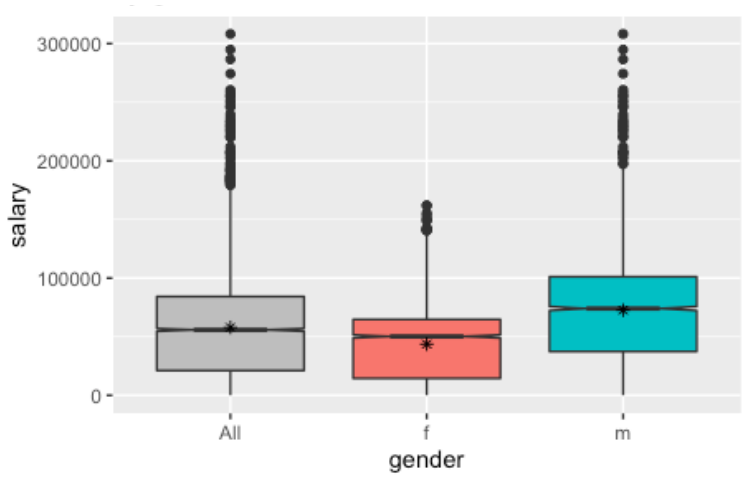
   Minimum is 2 and maximum is 99. The median is 46. Use the median to split sorted L to get {2, 5, 13, 15, 20, 32, 42, 42} and {55, 56, 59, 74, 76, 84, 92, 99}. Find the median of them as Q1 and Q3. Therefore, Q1 is 17.5 and Q2 is 75. The five number summary of L is {2, 17.5, 46, 75, 99}

Question 3

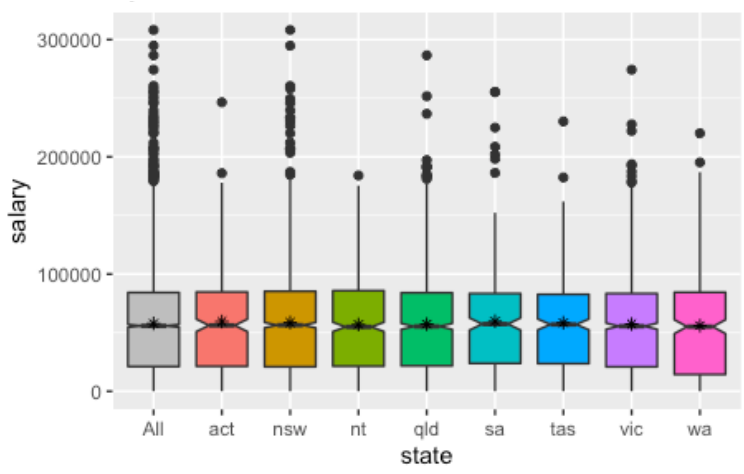
4. {14, 14, 14, 14, 14, 14}, {55, 55, 55, 55, 55}, {84, 84, 84, 84, 84}
5. {2, 2, 2, 2, 2, 42, 42, 42}, {55, 55, 55, 55, 55, 99, 99, 99}
6. {2, 2, 20, 20, 20}, {32, 42, 42}, {55, 55, 55, 74}, {76, 76, 99, 99}
7. {8.75, 8.75, 8.75, 8.75}, {34, 34, 34, 34}, {61, 61, 61, 61}, {87.75, 87.75, 87.75, 87.75}

Question 4

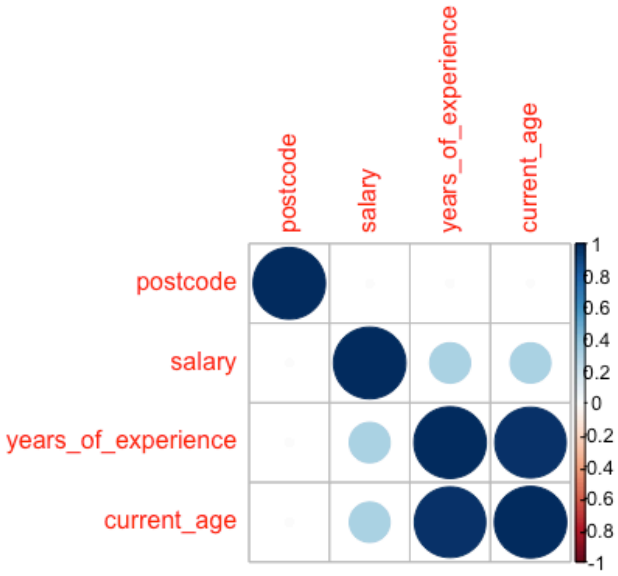
1. I want to find is there a gap between salary of male and female. As the figure shown below, the male's salary is higher than female's salary.



2. I want to find is there a gap between salary of different states. As the figure shown below, there is no obvious gap between those states' salary.



3. I want to find are salary and years of experience correlated. As the figure shown below, salary and years of experience are slightly correlated.



Question 5

	gender	postcode	state	phone	(age, birth_date)	marital_status
Completeness	100%	97.84%	97.84%	98.86%		98.72%
Consistency	100%	97.84%	97.84%	98.86%		95.6%
Accuracy					100%	

Completeness: the percentage of records that don't miss data in this attribute. For example, N/A is missing data

Consistency: the percentage of records which data follow this attribute's format. For example, for marital_status, NaN doesn't follow the format.

Accuracy: for (age, birth_date), it's the percentage of records which age add birth year should be this year.