MONASH University

Research Methods in IT

Week 8 Assessment
"Telling a data story"
Assessment Template

Student Name: Ziqi Pei
Student Number: 33429472
Tutor Name: : Pranita Shrestha

Task 1.

The data of this study included a total of 397 observed values of 4 variables. The data set was the income data of teachers in a certain university, including doctoral graduation time, working time in school, sex and salary.

graduation_time	working_time	sex	salary
19	18	Male	139750
20	16	Male	173200
4	3	Male	79750
45	39	Male	115000
40	41	Male	141500
6	6	Male	97000
30	23	Male	175000
45	45	Male	147765
21	20	Male	119250
18	18	Female	129000
12	8	Male	119800
7	2	Male	79800
1	1	Male	77700
2	0	Male	78000
20	18	Male	104800
12	3	Male	117150
19	20	Male	101000
38	34	Male	103450
37	23	Male	124750
39	36	Female	137000
31	26	Male	89565

36	31	Male	102580
34	30	Male	93904
24	19	Male	113068
13	8	Female	74830
21	8	Male	106294
35	23	Male	134885
5	3	Male	82379
11	0	Male	77000
12	8	Male	118223
20	4	Male	132261
7	2	Male	79916
13	9	Male	117256
4	2	Male	80225
4	2	Female	80225
5	0	Female	77000
22	21	Male	155750
7	4	Male	86373
41	31	Male	125196
9	9	Male	100938
23	2	Male	146500
23	23	Male	93418
40	27	Male	101299
38	38	Male	231545
19	19	Male	94384
25	15	Male	114778
40	28	Male	98193
23	19	Female	151768
25	25	Female	140096
1	1	Male	70768
28	28	Male	126621
12	11	Male	108875
11	3	Female	74692
16	9	Male	106639
12	11	Male	103760
14	5	Male	83900
23	21	Male	117704
9	8	Male	90215
10	9	Male	100135
8	3	Male	75044
9	8	Male	90304
3	2	Male	75243
33	31	Male	109785
11	11	Female	103613
4	3	Male	68404

9	8	Male	100522
22	12	Male	101000
35	31	Male	99418
17	17	Female	111512
28	36	Male	91412
17	2	Male	126320
45	45	Male	146856
29	19	Male	100131
35	34	Male	92391
28	23	Male	113398
8	3	Male	73266
17	3	Male	150480
26	19	Male	193000
3	1	Male	86100
6	2	Male	84240
43	28	Male	150743
17	16	Male	135585
22	20	Male	144640
6	2	Male	88825
17	18	Female	122960
15	14	Male	132825
37	37	Male	152708
2	2	Male	88400
25	25	Male	172272
9	7	Male	107008
10	5	Female	97032
10	7	Male	105128
10	7	Male	105631
38	38	Male	166024
21	20	Male	123683
4	0	Male	84000
17	12	Male	95611
13	7	Male	129676
30	14	Male	102235
41	26	Male	106689
42	25	Male	133217
28	23	Male	126933
16	5	Male	153303
20	14	Female	127512
18	10	Male	83850
31	28	Male	113543
11	8	Male	82099
10	8	Male	82600
15	8	Male	81500

40	31	Male	131205
20	16	Male	112429
19	16	Male	82100
3	1	Male	72500
37	37	Male	104279
12	0	Female	105000
21	9	Male	120806
30	29	Male	148500
39	36	Male	117515
4	1	Male	72500
5	3	Female	73500
14	14	Male	115313
32	32	Male	124309
24	22	Male	97262
25	22	Female	62884
24	22	Male	96614
54	49	Male	78162
28	26	Male	155500
2	0	Female	72500
32	30	Male	113278
4	2	Male	73000
11	9	Male	83001
56	57	Male	76840
10	8	Female	77500
3	1	Female	72500
35	25	Male	168635
20	18	Male	136000
16	14	Male	108262
17	14	Male	105668
10	7	Male	73877
21	18	Male	152664
14	8	Male	100102
15	10	Male	81500
19	11	Male	106608
3	3	Male	89942
27	27	Male	112696
28	28	Male	119015
4	4	Male	92000
27	27	Male	156938
36	26	Female	144651
4	3	Male	95079
14	12	Male	128148
4	4	Male	92000
21	9	Male	111168

12	10	Female	103994
4	0	Male	92000
21	21	Male	118971
12	18	Male	113341
1	0	Male	88000
6	6	Male	95408
15	16	Male	137167
2	2	Male	89516
26	19	Male	176500
22	7	Male	98510
3	3	Male	89942
1	0	Male	88795
21	8	Male	105890
16	16	Male	167284
18	19	Male	130664
8	6	Male	101210
25	18	Male	181257
5	5	Male	91227
19	19	Male	151575
37	24	Male	93164
20	20	Male	134185
17	6	Male	105000
28	25	Male	111751
10	7	Male	95436
13	9	Male	100944
27	14	Male	147349
3	3	Female	92000
11	11	Male	142467
18	5	Male	141136
8	8	Male	100000
26	22	Male	150000
23	23	Male	101000
33	30	Male	134000
13	10	Female	103750
18	10	Male	107500
28	28	Male	106300
25	19	Male	153750
22	9	Male	180000
43	22	Male	133700
19	18	Male	122100
19	19	Male	86250
48	53	Male	90000
9	7	Male	113600
4	4	Male	92700

4	4	Male	92000
34	33	Male	189409
38	22	Male	114500
4	4	Male	92700
40	40	Male	119700
28	17	Male	160400
17	17	Male	152500
19	5	Male	165000
21	2	Male	96545
35	33	Male	162200
18	18	Male	120000
7	2	Male	91300
20	20	Male	163200
4	3	Male	91000
39	39	Male	111350
15	7	Male	128400
26	19	Male	126200
11	1	Male	118700
16	11	Male	145350
15	11	Male	146000
29	22	Male	105350
14	7	Female	109650
13	11	Male	119500
21	21	Male	170000
23	10	Male	145200
13	6	Male	107150
34	20	Male	129600
38	35	Male	87800
20	20	Male	122400
3	1	Male	63900
9	7	Male	70000
16	11	Male	88175
39	38	Male	133900
29	27	Female	91000
26	24	Female	73300
38	19	Male	148750
36	19	Female	117555
8	3	Male	69700
28	17	Male	81700
25	25	Male	114000
7	6	Female	63100
46	40	Male	77202
19	6	Male	96200
5	3	Male	69200

31	30	Male	122875
38	37	Male	102600
23	23	Male	108200
19	23	Male	84273
17	11	Female	90450
30	23	Male	91100
21	18	Male	101100
28	23	Male	128800
29	7	Male	204000
39	39	Male	109000
20	8	Male	102000
31	12	Male	132000
4	2	Female	77500
28	7	Female	116450
12	8	Male	83000
22	22	Male	140300
30	23	Male	74000
9	3	Male	73800
32	30	Male	92550
41	33	Male	88600
45	45	Male	107550
31	26	Male	121200
31	31	Male	126000
37	35	Male	99000
36	30	Male	134800
43	43	Male	143940
14	10	Male	104350
47	44	Male	89650
13	7	Male	103700
42	40	Male	143250
42	18	Male	194800
4	1	Male	73000
8	4	Male	74000
8	3	Female	78500
12	6	Male	93000
52	48	Male	107200
31	27	Male	163200
24	18	Male	107100
46	46	Male	100600
39	38	Male	136500
37	27	Male	103600
51	51	Male	57800
45	43	Male	155865
8	6	Male	88650

49	49	Male	81800
28	27	Male	115800
2	0	Male	85000
29	27	Male	150500
8	5	Male	74000
33	7	Male	174500
32	28	Male	168500
39	9	Male	183800
11	1	Male	104800
19	7	Male	107300
40	36	Male	97150
18	18	Male	126300
17	11	Male	148800
49	43	Male	72300
45	39	Male	70700
39	36	Male	88600
27	16	Male	127100
28	13	Male	170500
14	4	Male	105260
46	44	Male	144050
33	31	Male	111350
7	4	Male	74500
31	28	Male	122500
5	0	Male	74000
22	15	Male	166800
20	7	Male	92050
14	9	Male	108100
29	19	Male	94350
35	35	Male	100351
22	6	Male	146800
6	3	Male	84716
12	9	Female	71065
46	45	Male	67559
16	16	Male	134550
16	15	Male	135027
24	23	Male	104428
9	9	Male	95642
13	11	Male	126431
24	15	Female	161101
30	31	Male	162221
8	4	Male	84500
23	15	Male	124714
37	37	Male	151650
10	10	Male	99247

23	23	Male	134778
49	60	Male	192253
20	9	Male	116518
18	10	Female	105450
33	19	Male	145098
19	6	Female	104542
36	38	Male	151445
35	23	Male	98053
13	12	Male	145000
32	25	Male	128464
37	15	Male	137317
13	11	Male	106231
17	17	Female	124312
38	38	Male	114596
31	31	Male	162150
32	35	Male	150376
15	10	Male	107986
41	27	Male	142023
39	33	Male	128250
4	3	Male	80139
27	28	Male	144309
56	49	Male	186960
38	38	Male	93519
26	27	Male	142500
22	20	Male	138000
8	1	Male	83600
25	21	Male	145028
49	40	Male	88709
39	35	Male	107309
28	14	Female	109954
11	4	Male	78785
14	11	Male	121946
23	15	Female	109646
30	30	Male	138771
20	17	Male	81285
43	43	Male	205500
43	40	Male	101036
15	10	Male	115435
10	1	Male	108413
35	30	Male	131950
33	31	Male	134690
13	8	Male	78182
23	20	Male	110515
12	7	Male	109707

30	26	Male	136660
27	19	Male	103275
28	26	Male	103649
4	1	Male	74856
6	3	Male	77081
38	38	Male	150680
11	8	Male	104121
8	3	Male	75996
27	23	Male	172505
8	5	Male	86895
44	44	Male	105000
27	21	Male	125192
15	9	Male	114330
29	27	Male	139219
29	15	Male	109305
38	36	Male	119450
33	18	Male	186023
40	19	Male	166605
30	19	Male	151292
33	30	Male	103106
31	19	Male	150564
42	25	Male	101738
25	15	Male	95329
8	4	Male	81035

Task 2: Craft a Research Inquiry

Research Question: How can we investigate the correlation between salary levels and the timing of doctoral graduation and develop a suitable model to analyze this relationship effectively?

Task3. For this data set, two key descriptive metric can be calculated:

Using two different methods, we explored the dataset's key metrics. The mean values for graduation time, working time, and salary are as follows:
Graduation Time: Mean = 22.31 years
Working Time: Mean = 17.61 years
Salary: Mean = $113,706

$$Mean = SunofValues / NumberofDatapoint$$

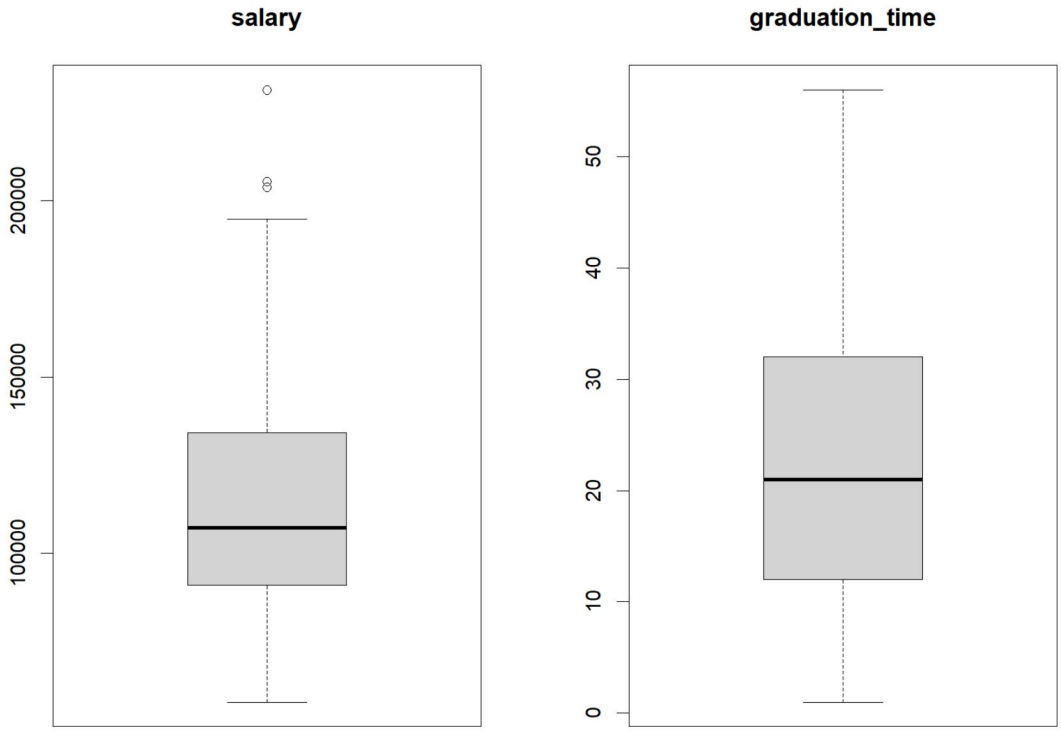
Additionally, I performed Multiple linear regression analysis was used to analyze the data, and the final equation was as follows:

$$Salary = 89912.2 + 985.3 * garduation\_time$$

Task4. Appropriate Visualization

I employed data visualization in R using the following code. The boxplot below illustrates the distribution of two variables. It is evident from the plot that only the salary variable contains several outliers, while the graduation_time variable is free from outliers.

```r
#Load package
library(openxlsx)
library(ggplot2)
library(ggfortify)
#Data read
data <- read.xlsx('data.xlsx')
#Statistical description (mean, median, etc.)
summary(data)
#Model construction
res <- lm(data = data,salary~graduation_time)
#View model summary
summary(res)
#Draw the box diagram
par(mfrow = c(1,2))
boxplot(data[c('salary')],main = 'salary')
boxplot(data[c('graduation_time')],main = 'graduation_time')
#Model diagnosis
par(mfrow = c(2,2))
autoplot(lm(data = data,salary~graduation_time))
```



The diagnostic graph of the regression model is as follows:

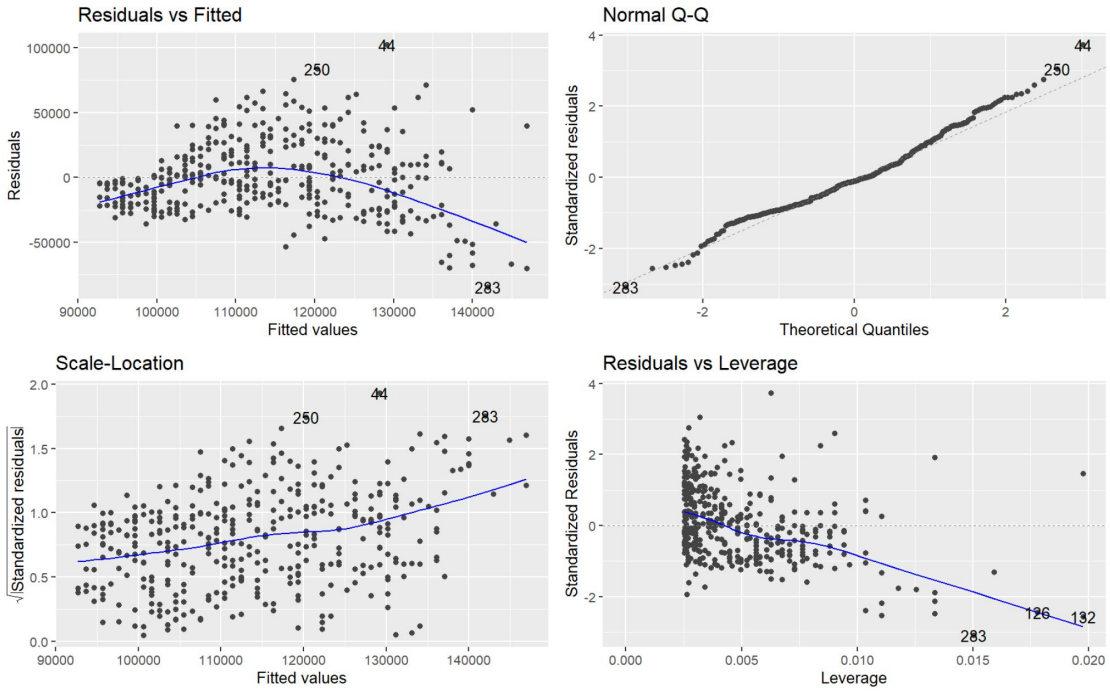
The diagnostic graph of the regression model provides a visual assessment of how well the model fits the data and whether the model assumptions are met. It typically includes various plots to examine key aspects of the regression analysis.

Residual vs. Fitted: This plot helps us understand if there is a pattern in the residuals (the differences between observed and predicted values). A random scatter of residuals around zero indicates that the model is appropriate.

Normal Q-Q (Quantile-Quantile): This plot assesses the normality of the residuals. If the residuals closely follow a diagonal line, it suggests that the residuals are normally distributed.

Scale-Location (Spread-Location): This plot examines whether the spread of residuals is consistent across the range of fitted values. Ideally, the points should form a horizontal band with no clear funnel shape.

Residual vs. Leverage: This plot identifies influential data points that may have a significant impact on the regression model. Data points that fall outside a certain threshold can be considered influential.



Task5 .Description finding

The regression equation shows that the average salary of teachers increases with the increase of the years of doctoral graduation. On average, the average salary increases by 985.3 for each additional year of doctoral graduation. The variance analysis of the whole regression model obtained F value of 84.23, P<0.05, the difference was statistically significant, so the model was effective. The diagnostic graph of the model shows that the data meet the conditions of normality and homogeneity of variance. The box diagram shows that there is a small amount of outlier value in salary, but it does not have a large impact on the analysis.