STATISTICAL DATA ANALYSIS COURSE USING R PROGRAMMING LANGUAGE: A VIRTUAL TRAINING

MARCH 21st –26th, 2022

Exploratory data analysis

Assoc.Prof.Susan Balaba Tumwebaze
Dr.Thomas L Odong
Dr.Hellen Namawejje

What is Exploratory Data Analysis

- Data analysis aimed at summarizing the main characteristics of data sets using numerical and graphical methods
- A critical first step in analyzing data which provide insight before formal modeling and hypothesis testing
- "Promoted by John Tukey to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments"(**thereaderwiki.com/en/Exploratory_data_analysis).**

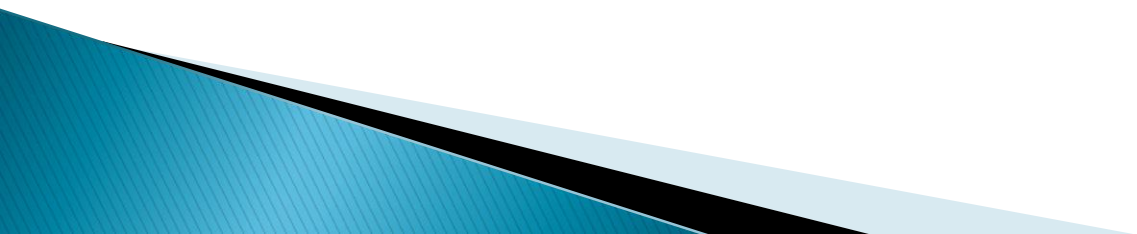
Importance of Exploratory Data Analysis

- Maximize insight into a data set;
- Uncover underlying structure - including relationships between variables;
- Extract important variables – including creating new variables and data transformation
- Detection of outliers and mistakes;

Importance of Exploratory Data Analysis

- Test underlying assumptions on which statistical inference will be based;
- Suggest hypotheses about the causes of observed phenomena
- Support the selection of appropriate statistical tools and techniques
- Provide a basis for further data collection through surveys or experiments

Overview of Methods for Exploratory Data Analysis



Categorization of EDA methods

- The choice of methods for exploratory data analysis will depend on;

1) Method of summarizing the data
   - Numerical (Non-graphical) versus Graphical methods

2) Number of variables being examined at a time
   - Univariate (one variable at a time) versus Multivariate (two or more variables at a time)

Categorization of EDA methods

	Non-graphical	Graphical
Univariate	• Measures of locations (mean, median, mode, quartiles) • Measures of spread (range, variance, standard deviations, inter-quartile range) • Frequency tables	• Box plot • Histogram, stem-and-leaf plot • Bar graph • Pie chart
Multivariate	• Cross-tabulation • Covariance and correlation coefficient • Variance-covariance matrix	• Scatter plot • Bar graph, histogram • Box plot

The four types of EDA are univariate non-graphical, multivariate non-graphical, univariate graphical, and multivariate graphical.

Type of Variables

- The choice of summary statistics also depends on the types of variables we are dealing with.
- Two major types of variables: ***qualitative*** (categorical) and ***quantitative***

Quantitative variables

Discrete if the measurements are integers, obtained by counting(e.g. number of people in a household, number of cigarettes smoked per day).

- **Continuous** if the measurements can take on any value, usually within some range (e.g. weight) and are obtained by measuring.

Types of Variables

Qualitative variables

- Describes a characteristics or attribute of an individual.
- It includes data that can be categorised but not quantified.
- Qualitative variables may be either ordinal or Nominal/ Categorical.

Nominal variables are associated with some quality, characteristic or attribute which the variable posses; e.g. eye colour, vehicle type.

Ordinal variables deal with relative differences, e.g., (short – average – tall), (bad – good – excellent), etc

Description of datasets used in the training

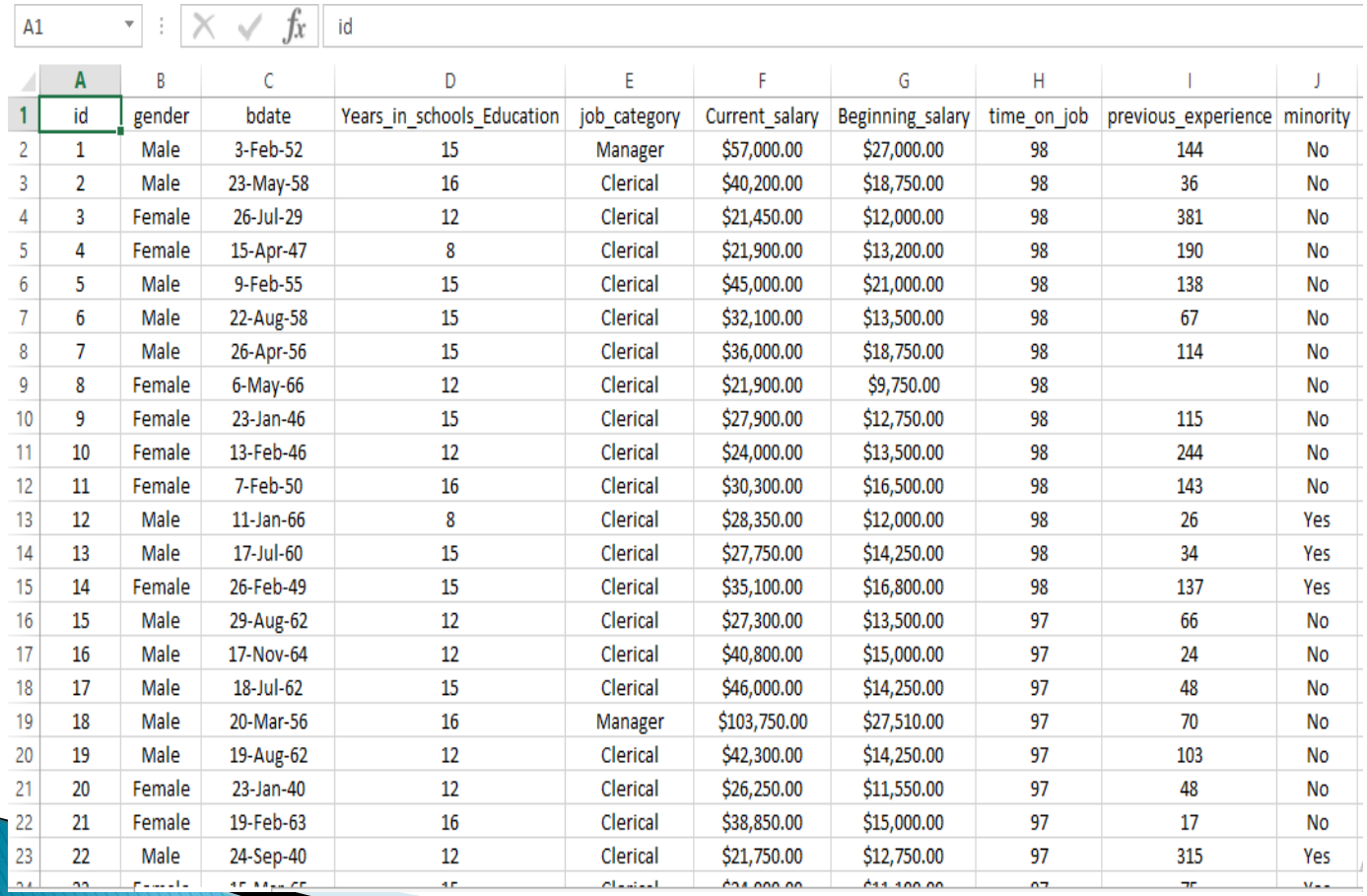
Description of datasets

Data : Employee data

- **Variables of interest**
  - Gender of employee (male or female)
  - Education level – number of years at school
  - Job categories (clerical, custodial, manager)
  - Current salary (US$)
  - Beginning salary(US$)
  - Time on job (months)
  - Previous experience (years)
  - Minority categorization (Yes or No)

Description of datasets

Data : Employee data



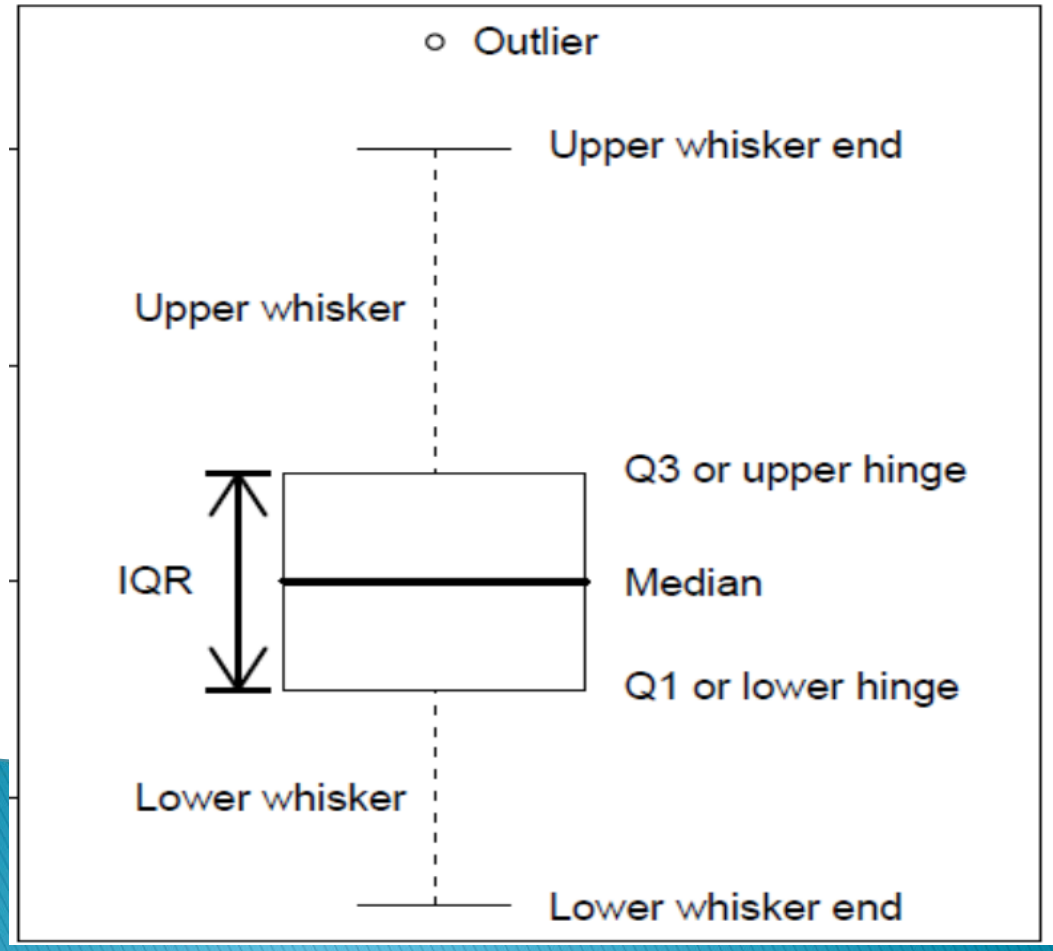
A1 | id

	A	B	C	D	E	F	G	H	I	J
1	id	gender	bdate	Years_in_schools_Education	job_category	Current_salary	Beginning_salary	time_on_job	previous_experience	minority
2	1	Male	3-Feb-52	15	Manager	$57,000.00	$27,000.00	98	144	No
3	2	Male	23-May-58	16	Clerical	$40,200.00	$18,750.00	98	36	No
4	3	Female	26-Jul-29	12	Clerical	$21,450.00	$12,000.00	98	381	No
5	4	Female	15-Apr-47	8	Clerical	$21,900.00	$13,200.00	98	190	No
6	5	Male	9-Feb-55	15	Clerical	$45,000.00	$21,000.00	98	138	No
7	6	Male	22-Aug-58	15	Clerical	$32,100.00	$13,500.00	98	67	No
8	7	Male	26-Apr-56	15	Clerical	$36,000.00	$18,750.00	98	114	No
9	8	Female	6-May-66	12	Clerical	$21,900.00	$9,750.00	98		No
10	9	Female	23-Jan-46	15	Clerical	$27,900.00	$12,750.00	98	115	No
11	10	Female	13-Feb-46	12	Clerical	$24,000.00	$13,500.00	98	244	No
12	11	Female	7-Feb-50	16	Clerical	$30,300.00	$16,500.00	98	143	No
13	12	Male	11-Jan-66	8	Clerical	$28,350.00	$12,000.00	98	26	Yes
14	13	Male	17-Jul-60	15	Clerical	$27,750.00	$14,250.00	98	34	Yes
15	14	Female	26-Feb-49	15	Clerical	$35,100.00	$16,800.00	98	137	Yes
16	15	Male	29-Aug-62	12	Clerical	$27,300.00	$13,500.00	97	66	No
17	16	Male	17-Nov-64	12	Clerical	$40,800.00	$15,000.00	97	24	No
18	17	Male	18-Jul-62	15	Clerical	$46,000.00	$14,250.00	97	48	No
19	18	Male	20-Mar-56	16	Manager	$103,750.00	$27,510.00	97	70	No
20	19	Male	19-Aug-62	12	Clerical	$42,300.00	$14,250.00	97	103	No
21	20	Female	23-Jan-40	12	Clerical	$26,250.00	$11,550.00	97	48	No
22	21	Female	19-Feb-63	16	Clerical	$38,850.00	$15,000.00	97	17	No
23	22	Male	24-Sep-40	12	Clerical	$21,750.00	$12,750.00	97	315	Yes

Summarizing Quantitative variable
(One variable at a time)

Five Number Summary and a Boxplot

BOXPLOT



Five number summary

Employee data

- **Which of the variables are quantitative ?**
  - Gender of employee (male or female)
  - Education level – number of years at school
  - Job categories (clerical, custodial, manager)
  - Current salary (US$)
  - Beginning salary(US$)
  - Time on job (months)
  - Previous experience (years)
  - Minority categorization (Yes or No)

Five number summary

Employee data

- **Quantitative variables**
  - Gender of employee (male or female)
  - **Education level – number of years at school**
  - Job categories (clerical, custodial, manager)
  - **Current salary (US$)**
  - **Beginning salary(US$)**
  - **Time on job (months)**
  - **Previous experience (years)**
  - Minority categorization (Yes or No)

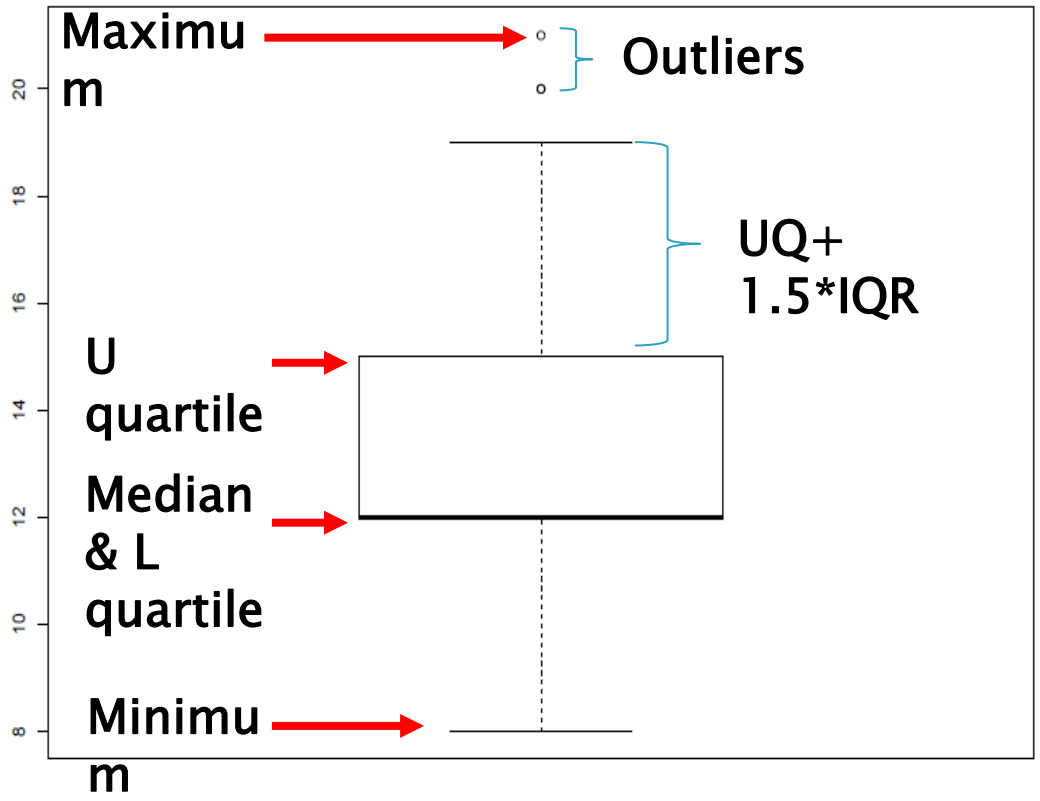
Five number summary

- **Variable**: Education level

Statistic (summary)	Years at school
Minimum	8
Lower quartile	12
Median	12
Upper Quartile	15
Maximum	21

- Least educated spent 8 years at school
- Most Educated spent 21 years at school
- 75% spent less than 15 years at school

Boxplot for level of Education



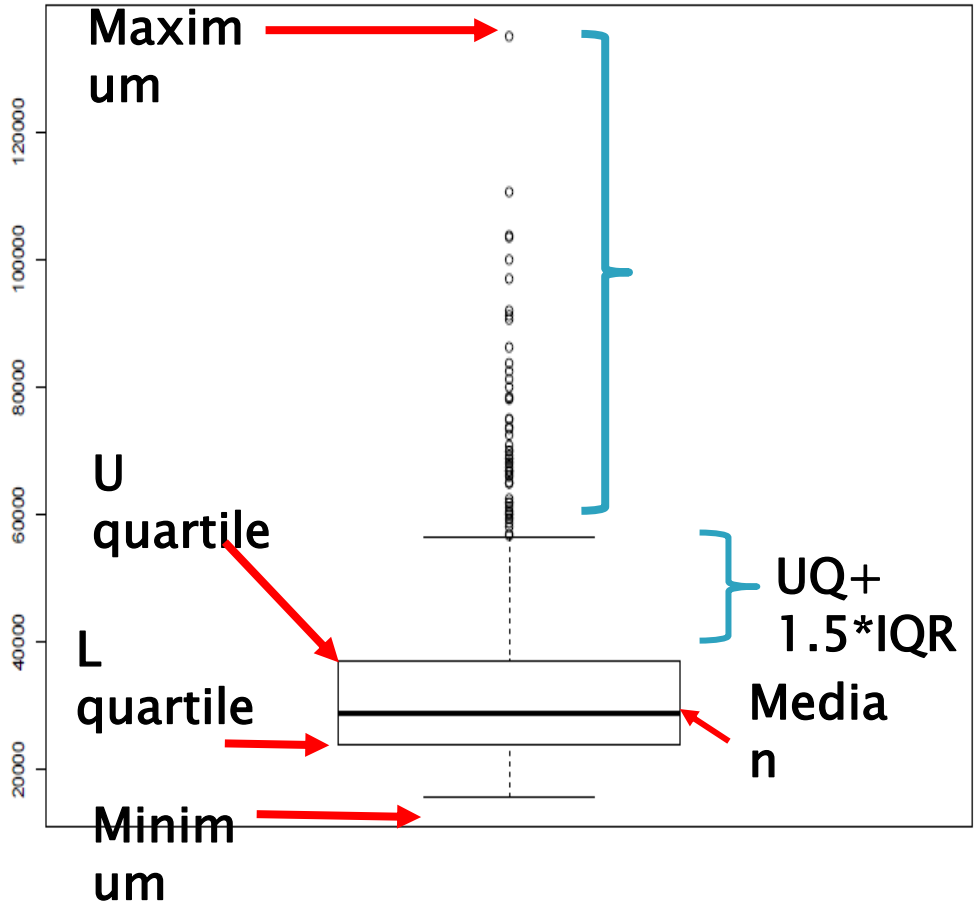
Interquartile range (IQR)=Upper quartile−Lower quartile

Boxplot for current salary

Five number summary

- **Variable:** Current salary

Statistic (summary)	Salary (US$)
Minimum	15,750
Lower quartile	24,000
Median	28,875
Upper Quartile	37,050
Maximum	135,000



- **Least paid earn US$15,750**
- **Highest paid earn US$135,000**
- **25% earn less than US$24,000**
- **75% earn below US$37,050**

Outliers

- Observations separated from the rest – unusual values
  - You need to explore the reason for unusual observations
  - Outlier could be of interest **if it is not an error**
  - Practice is often to do data analysis with and without the outliers and compare the results

Description of population distribution

- Population distribution characteristics of interest
  - Center
  - Spread
  - Shape (including “heaviness of the tails”)
  - Outliers

Description of population distribution

Measure of central tendency

Mean

- Commonly use measure center of the distribution

$$\bar{X} = \frac{\sum x_i}{n} = (x1 + x2 + \cdots . xn)/n$$

- Affected by extreme observations
- **Use of trimmed mean** – excluding extreme observation from calculated the mean

Description of population distribution

Measure of central tendency - Mean

Difference between the mean and median is an indication of Effect of extreme values on the means

Measure center	Current salary	Years at school
Mean	34,420.00	13.49
Median	28,800.00	12.00
trimmed mean	28,866.88 (80% off)	12.40 (92% off)

Description of population distribution

Measure of central tendency - Mean

Difference between the mean and median is an indication of **effect of extreme values** on the means

Percent trimmed	Mean current salary(US$)	Median current salary (US$)
0	34,420.00	28,800
5	32,438.09	28,875
10	30,976.53	28,800
15	30,115.18	28,800
20	29,540.32	28,800
25	29174.43	28,800

The median as a measure of center is very robust – does not change much with changes in data

Description of population distribution

Measures of spread

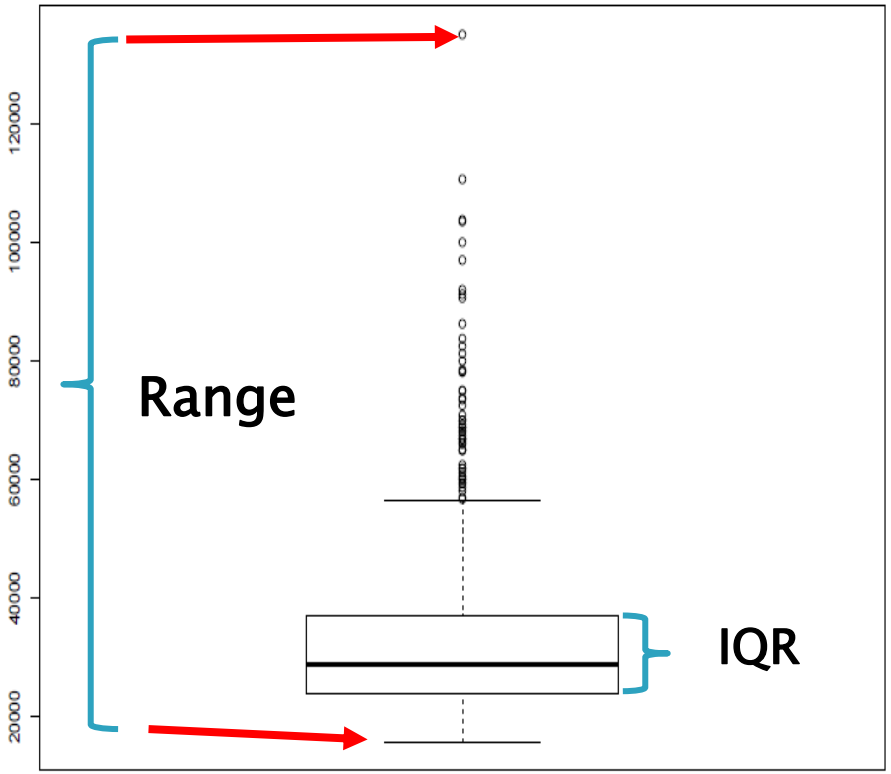
- Spread is a measure of how far a value is from the central value
- Measures of spread include
- Range=

Maximum – Minimum

- Interquartile range (IQR)

=Upper Quartile – Lower quartile

- Both Range and Interquartile range **do not** take into account all the data values



Boxplot for current salary

Description of population distribution

Measures of spread

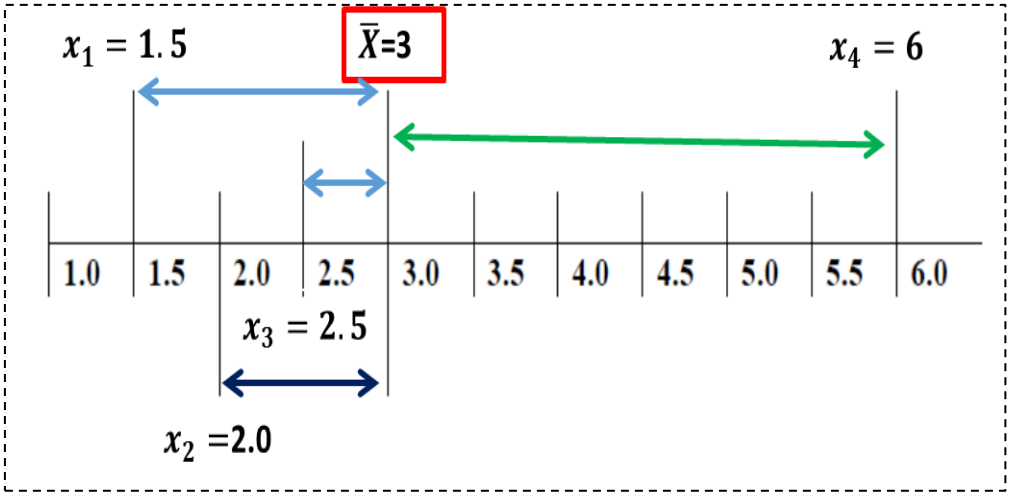
Variance and Standard deviation

- Based on squared deviation from the center (mean)

$$S^2 = \frac{\sum_{i=1}^{4}(x_i - \bar{x})^2}{(n-1)}$$

$$= \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + (x_4 - \bar{x})^2}{(n-1)}$$

$$= \frac{(1.5 - 3)^2 + (2.0 - 3)^2 + (2.5 - 3)^2 + (6 - 3)^2}{(4-1)}$$



- Standard deviation = Square root of the variance

$$SD = \sqrt{\frac{\sum_{i=1}^{4}(x_i - \bar{x})^2}{(n-1)}}$$

Description of population distribution

Measures of spread

Effect of removing extreme observations (trimming off) on the different measures of spread – **Years in school from the employee data**

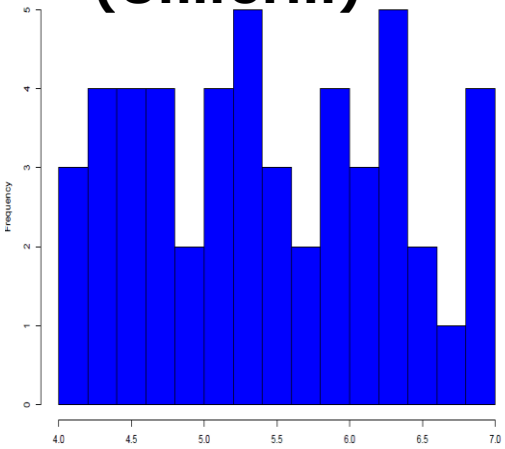
Percentage of data trimmed off	Range	Interquartile Range (IQR)	Variance	Stand deviation
0	13	3	8.322	2.885
5	11	3	5.728	2.393
10	8	3	3.298	1.816
15	4	3	2.668	1.633
20	4	3	2.411	1.553
25	3	3	2.197	1.482
Robustness	No	Yes	No	No

Description of population distribution

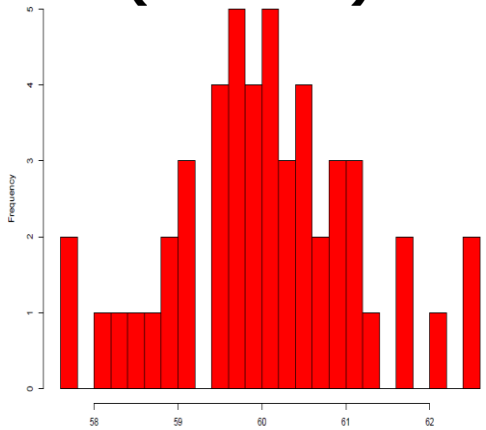
The shape of the distribution

- Histogram vary by
  - Whether the histogram has a peak or not
  - Number of peaks
  - How symmetric is it
  - How long each tail is in relation to the other

No peak (Uniform)



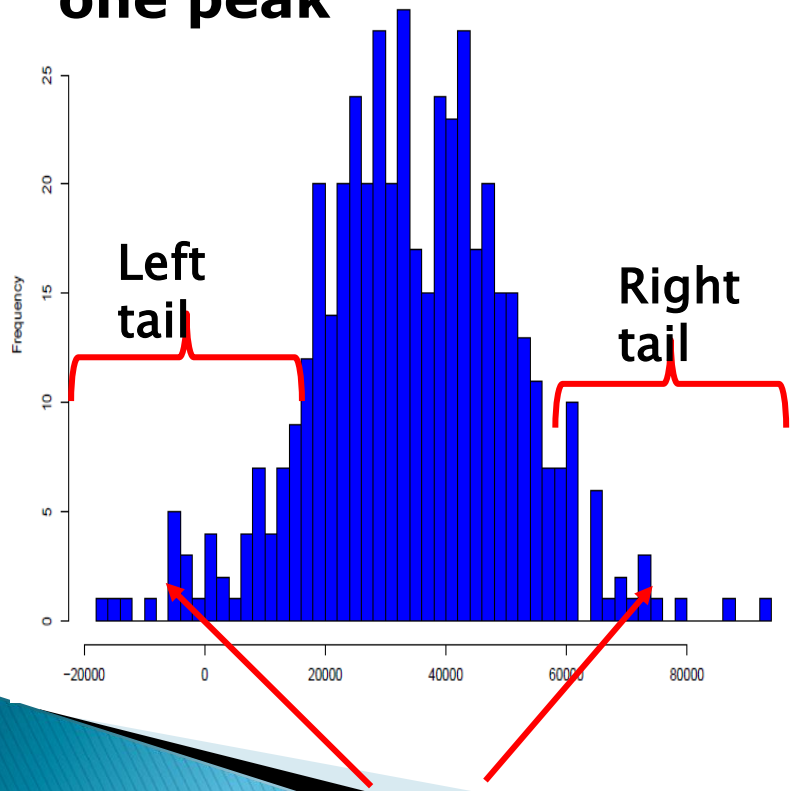
One peak (Normal)



Description of population distribution

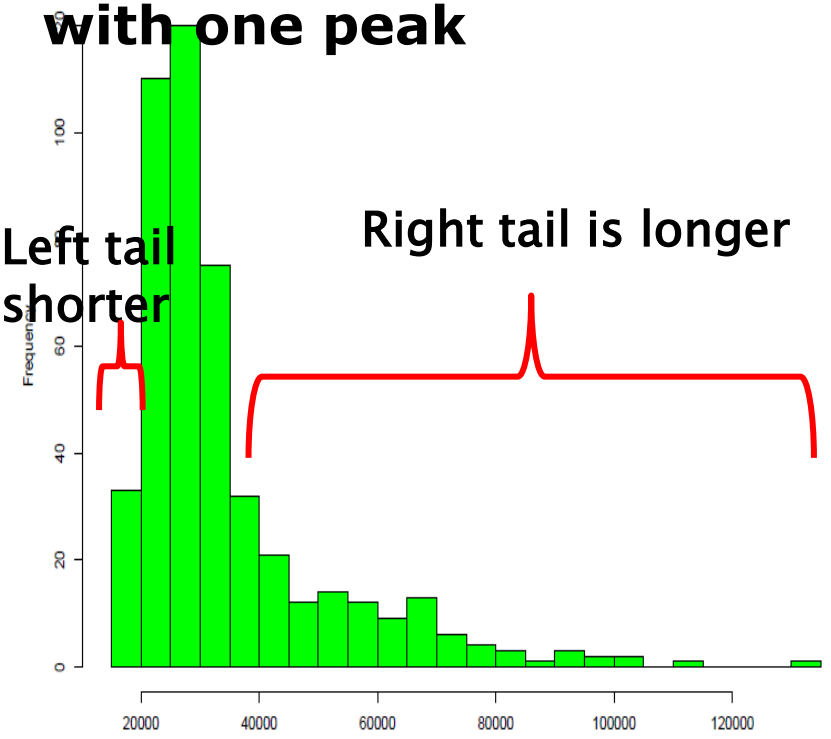
The shape of the distribution

Symmetric with one peak



Both tails are of similar length

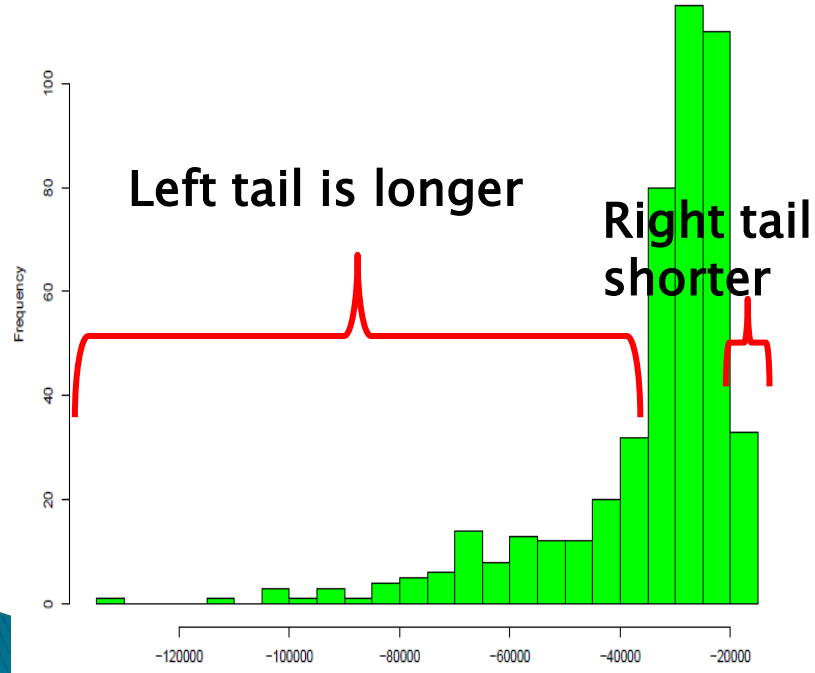
Non-symmetric with one peak



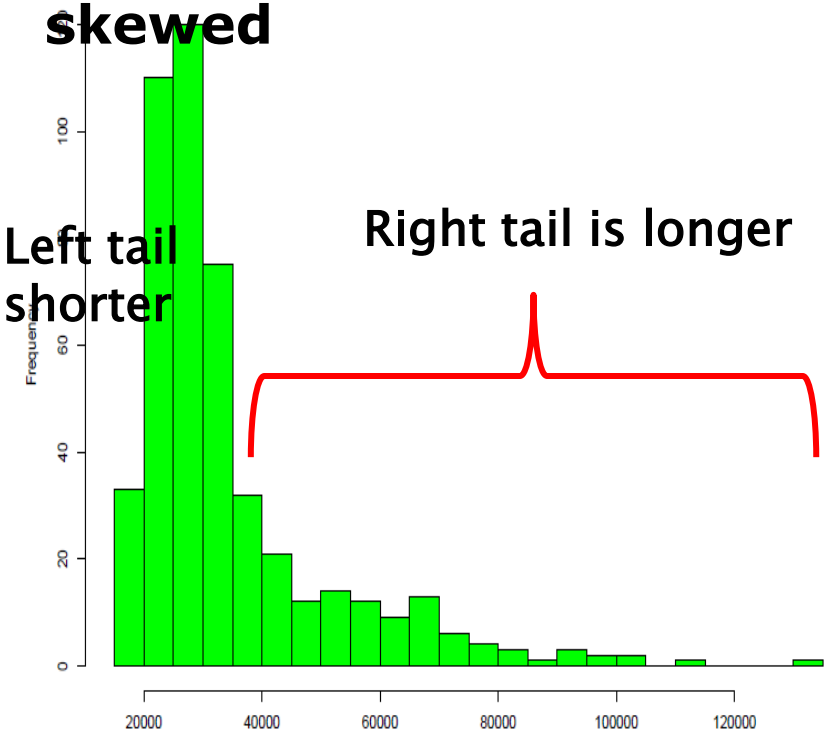
Description of population distribution

The shape of the distribution

Left or negative skewed



Right or positive skewed

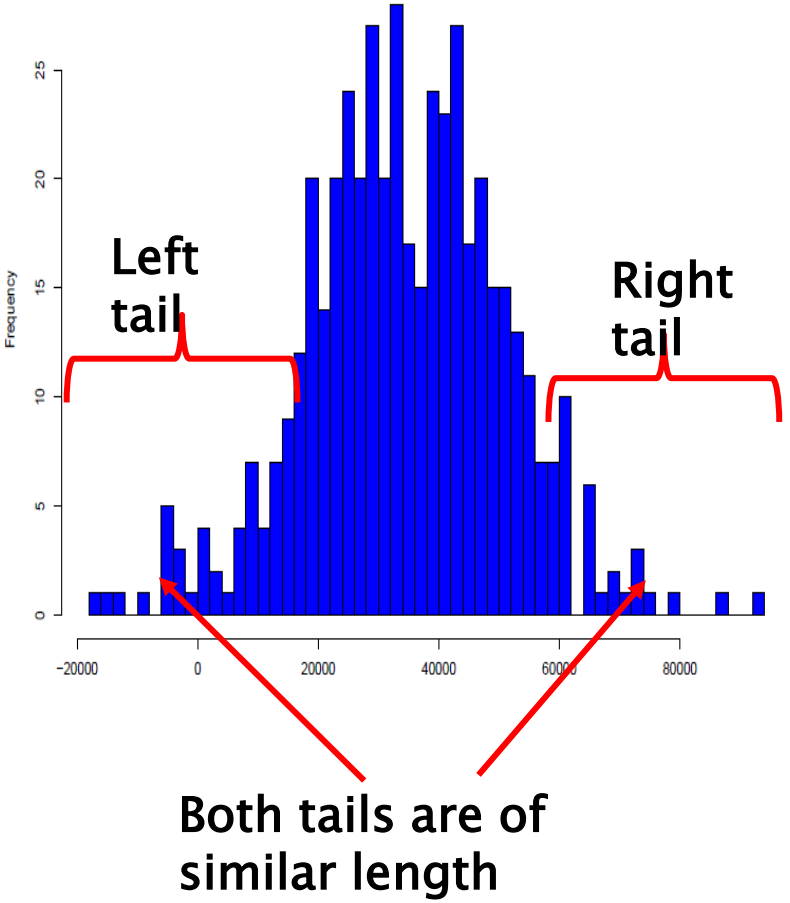


Description of population distribution

Symmetric distribution

- Median ~ Mean

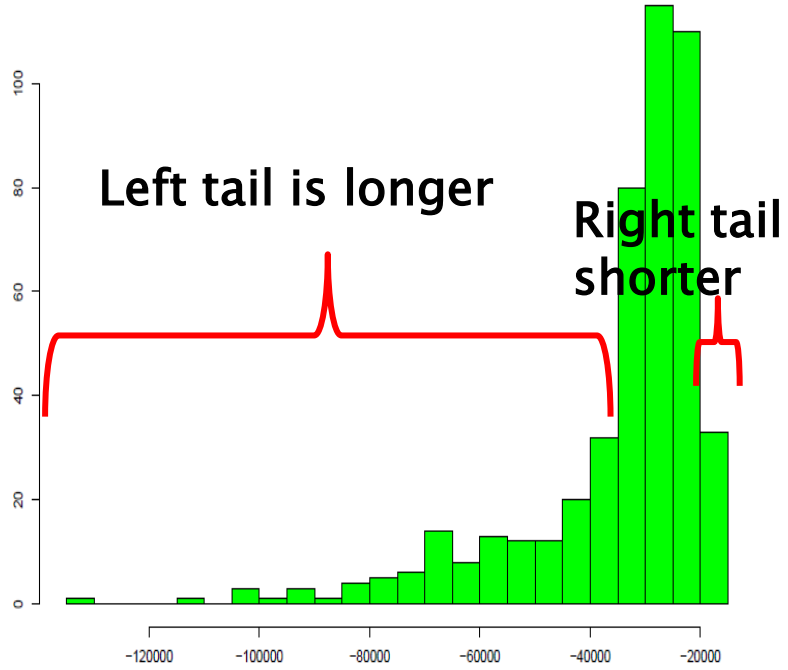
Minimum	**-17700**	
1st Q	22430	
Median	33700	
3rd Q	44,730	
Mean	33,940	
Maximum	83,580	



Description of population distribution

Minimum	-135,000	
1st Q	-36,940	Median- 1st Q = 8060
Median	-28,880	
3rd Q	-24,000	3rd Q - Median = 4880
Mean	-34,420	
Maximum	-15,750	

Left or negative skewed



Description of population distribution

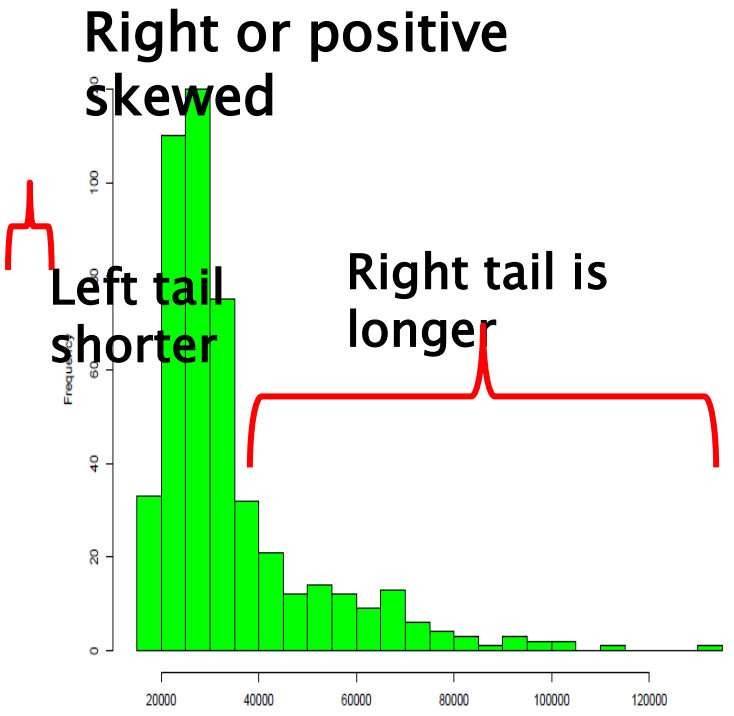
The shape of the distribution

Minimum	15,750	
1st Q	24,000	(Median– 1st Q) = 4880
Median	28,880	
3rd Q	36,940	(3rd Q – Median) = 8060
Mean	34,420	
Maximum	135,000	

Right or positive skewed

Mean > Median

(Median – 1st Q) < (3rd Q – Median)



Summarizing qualitative variables

Description of population distribution

- The concepts of central tendency, spread and skew have no meaning for qualitative nominal variables.
- For qualitative ordinal variables, it sometimes makes sense to treat the data as quantitative for EDA purposes; you need to use your judgment here.

Summarizing qualitative variables

- With quantitative variables we were able determine
  - Minimum, maximum observations
  - Median, mean, quartiles
  - Range, interquartile range, variance and standard deviation
  - Boxplot, histograms
  - Skewedness
  - Identification outliers

Summarizing qualitative variables

For Employee data, which variables are qualitative?

- **Qualitative variables**
  - **Gender of employee (male or female)**
  - Education level – number of years at school
  - **Job categories (clerical, custodial, manager)**
  - Current salary (US$)
  - Beginning salary(US$)
  - Time on job (months)
  - Previous experience (years)
  - **Minority categorization (Yes or No)**

Summarizing qualitative variables

- Frequency tables
- Pie chart
- Bar graph

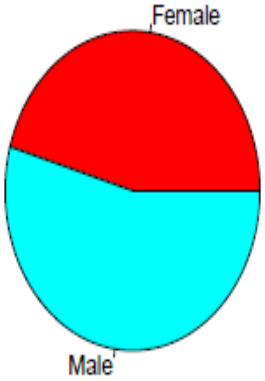
Summarizing qualitative variables

Frequency table

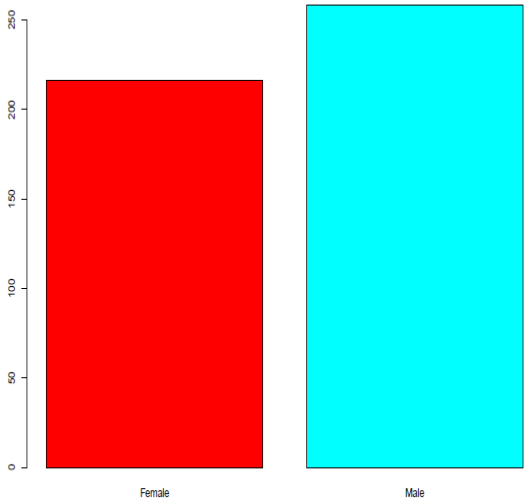
Gender	Frequency	Relative frequency
Female	216	0.456
Male	258	0.544
Total	474	1.000

Pie charts are a very bad way of displaying information. The eye is good at judging linear measures and bad at judging relative areas. A bar chart or dot chart is a preferable way of displaying this type of data

Pie chart



45.6% of the employee were female



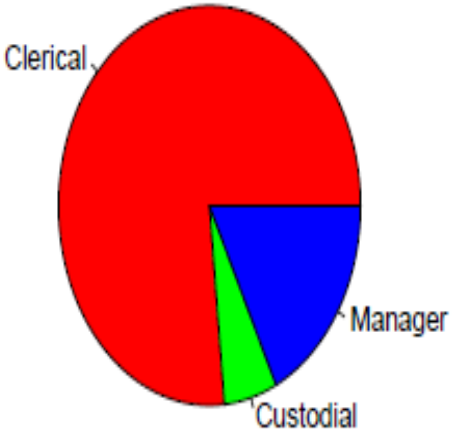
Bar graph

Summarizing qualitative variables

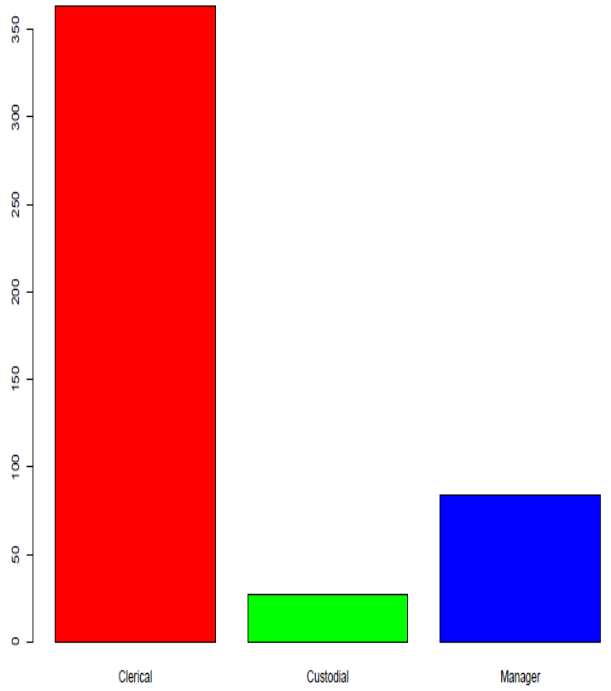
Frequency table

Job Category	Frequency	Relative frequency
Clerical	363	0.766
Custodial	27	0.057
Manager	84	0.177
Total	474	1.000

The majority (76.6%) of the employee were clerical staff



Pie chart



Bar graph

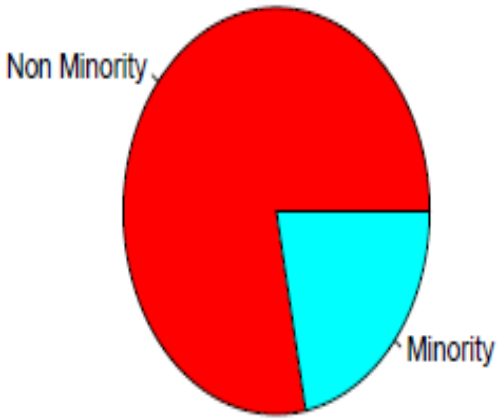
Summarizing qualitative variables

Frequency table

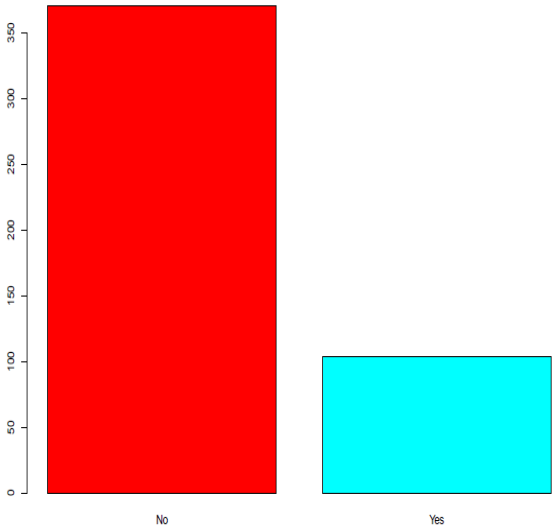
Minority Classification	Frequency	Relative frequency
Yes	104	0.219
No	370	0.781
Total	474	1.000

Pie charts are a very bad way of displaying information. The eye is good at judging linear measures and bad at judging relative areas. A bar chart or dot chart is a preferable way of displaying this

Pie chart

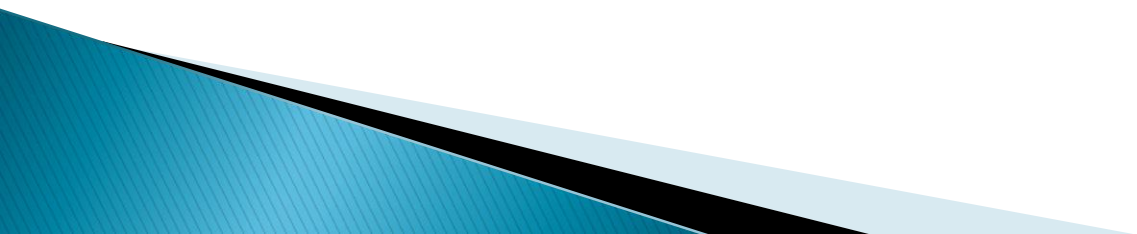


Bar graph



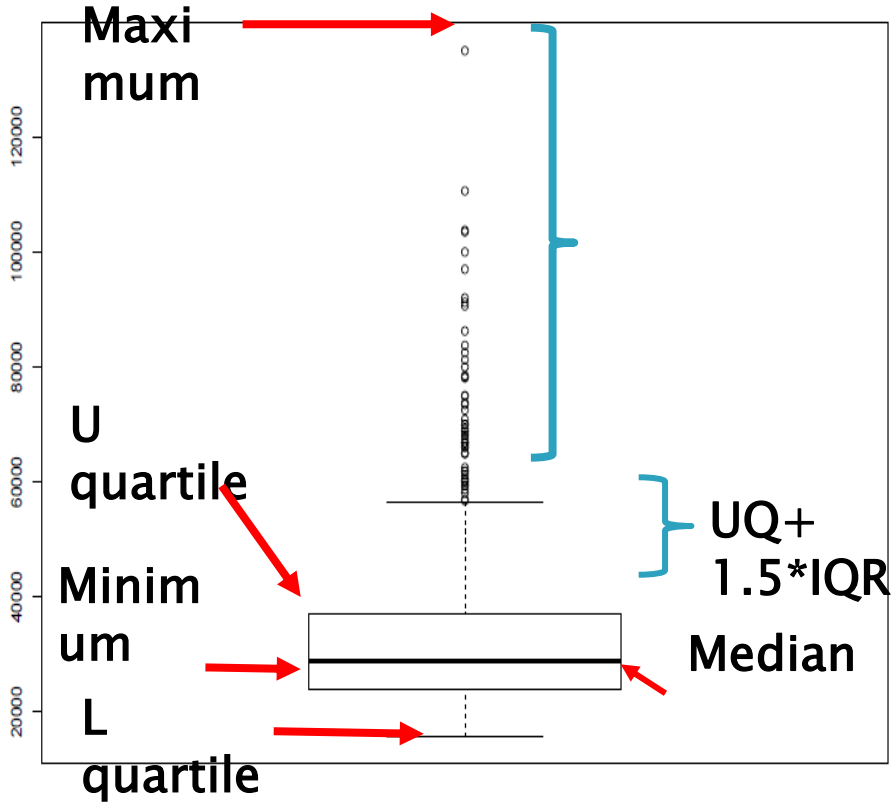
About 20% of the employee belong to minority grouping

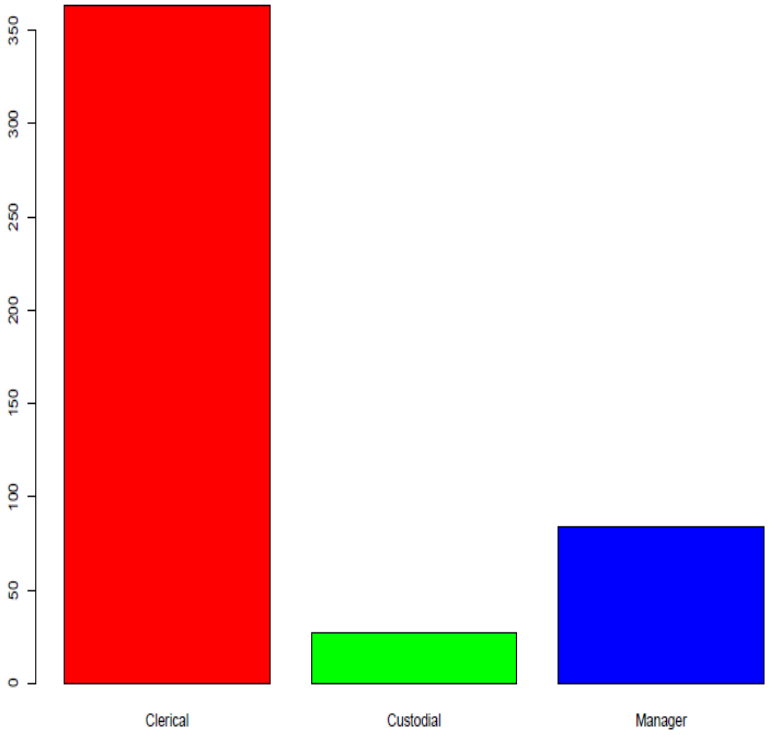
Exploring relationships between two variables



Relationships between qualitative and quantitative variables

Boxplot for current salary



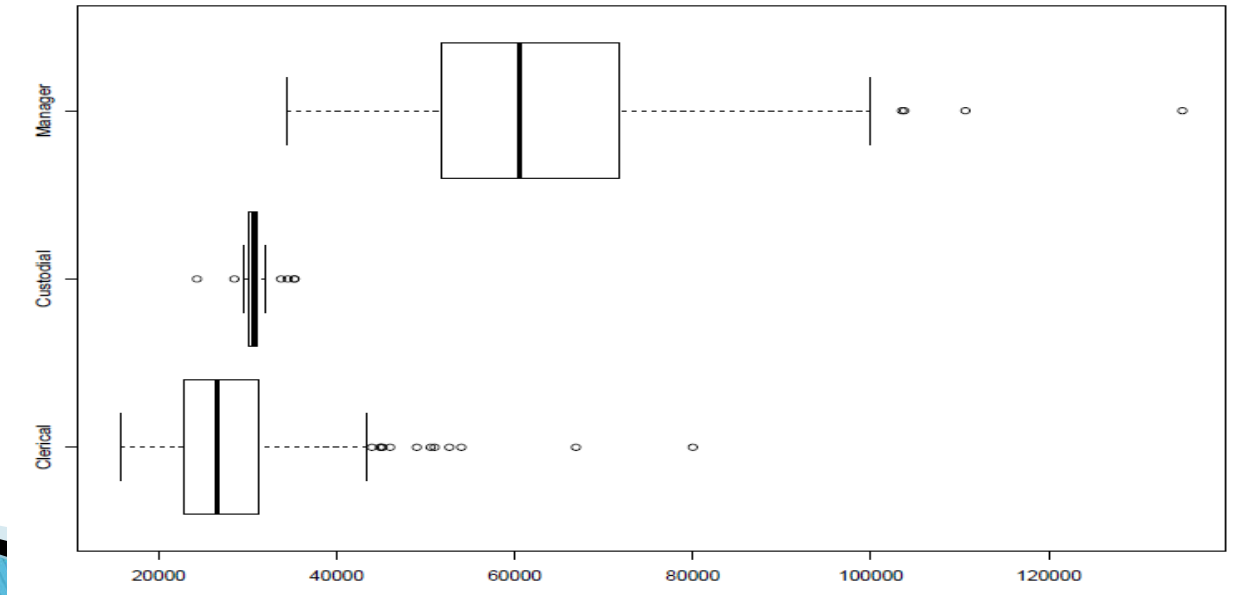


Bar graph for employ category

Is there relationship between current salary and employment category?

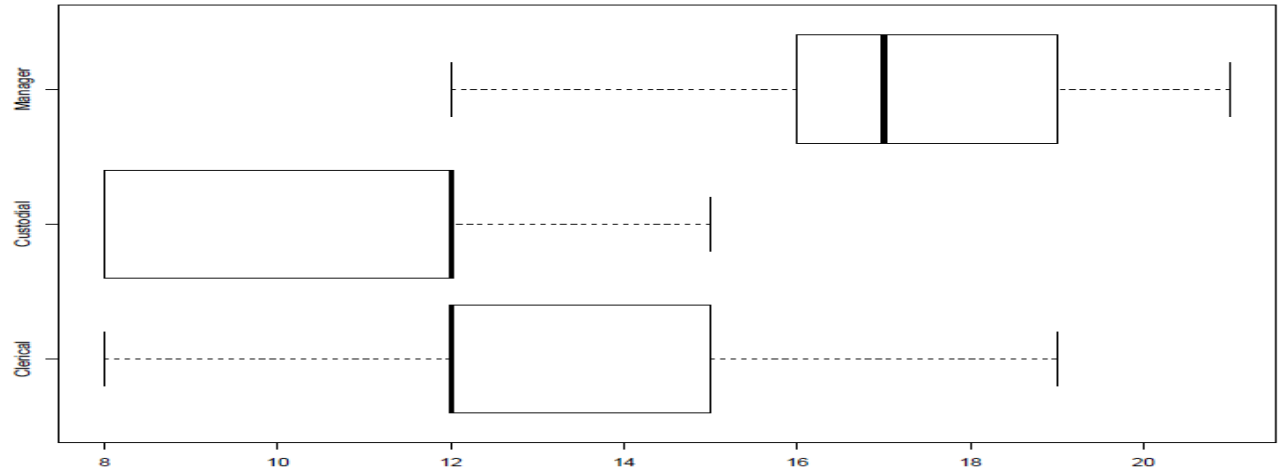
Relationships between qualitative and quantitative variables

	Summary Statistics for current salary by Job category					
Job category	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Clerical	15750	22800	26550	27840	31200	80000
Custodial	24300	30150	30750	30940	30980	35250
Manager	34410	51960	60500	63980	71280	135000



Relationships between qualitative and quantitative variables

	Summary of Education levels (Years at Schools)					
Job category	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Clerical	8	12	12	12.87	15	19
Custodial	8	8	12	10.19	12	15
Manager	12	16	17	17.25	19	21



Relationships between qualitative and quantitative variables

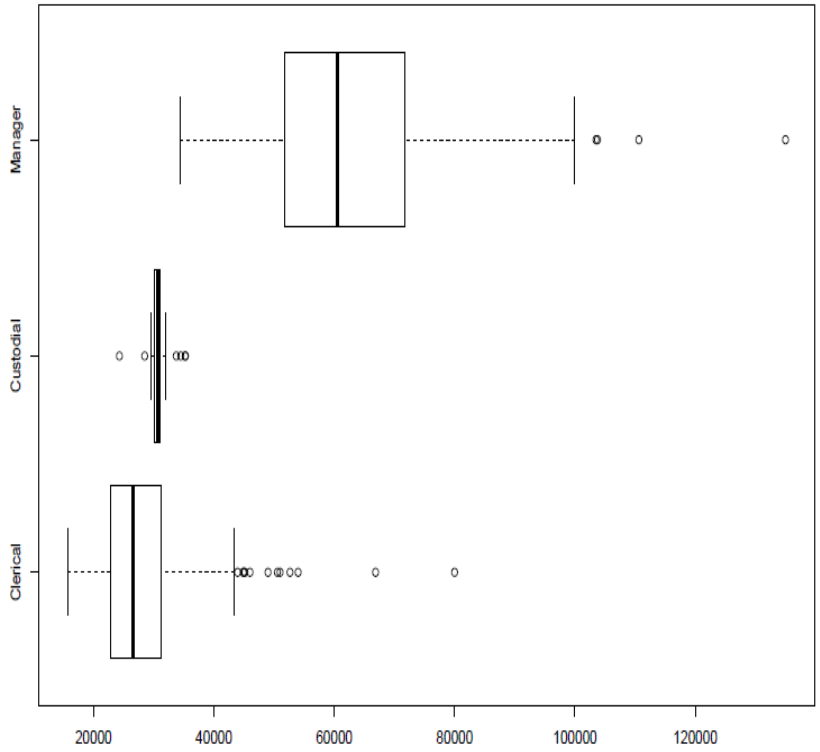
	Summary Statistics for current salary by Job category					
Job category	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Clerical	15750	22800	26550	27840	31200	80000
Custodial	24300	30150	30750	30940	30980	35250
Manager	34410	51960	60500	63980	71280	135000

	Summary of Education levels (Years at Schools)					
Job category	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Clerical	8	12	12	12.87	15	19
Custodial	8	8	12	10.19	12	15
Manager	12	16	17	17.25	19	21

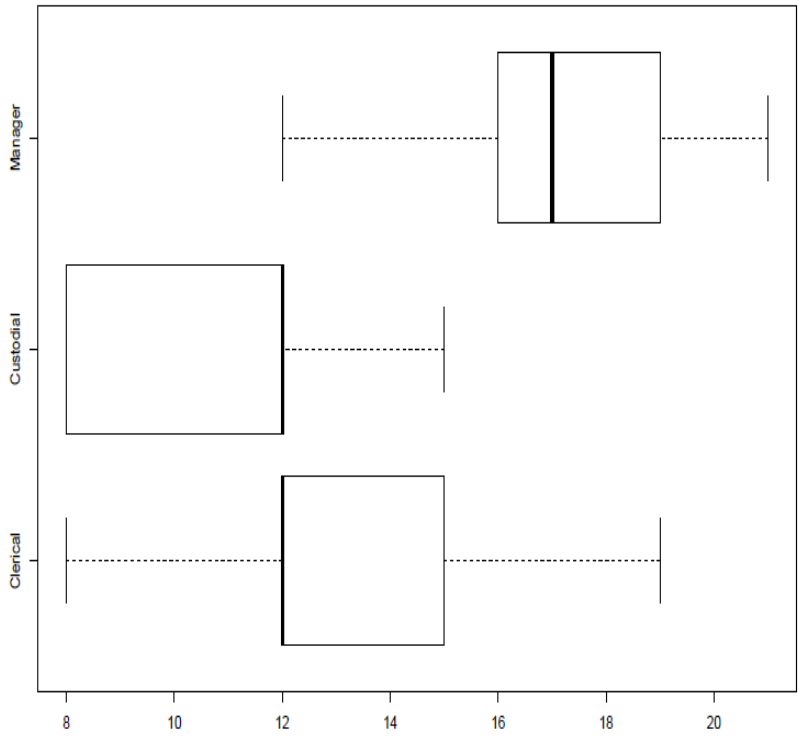
Can education level explain disparities in salary?

Relationships between qualitative and quantitative variables

Current Salary



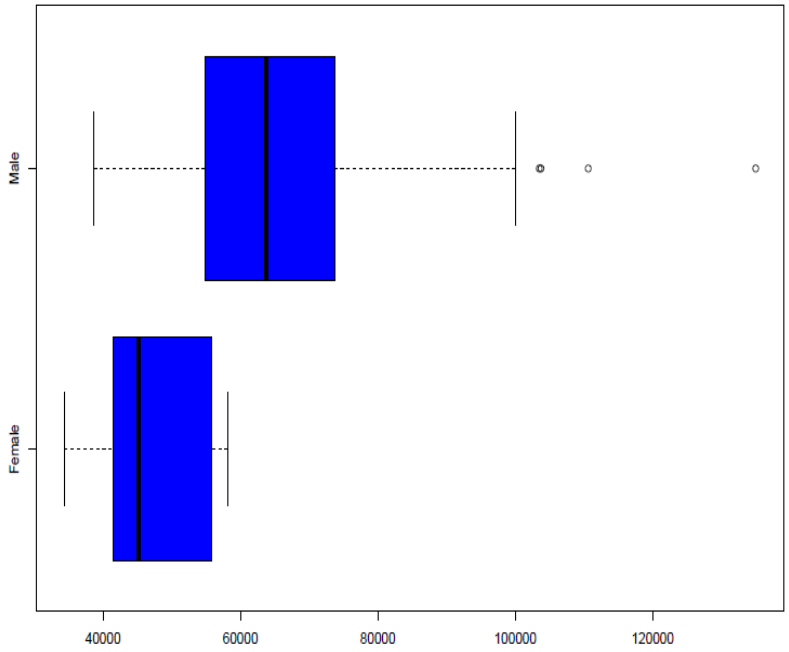
Education (Years at School)



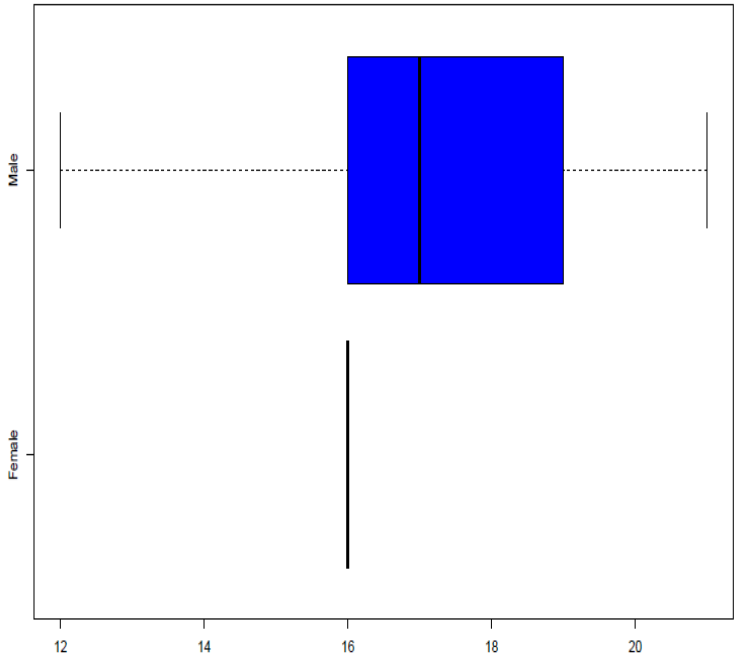
Can education explain disparities in salary with each category?

Relationships between qualitative and quantitative variables

Current salary of managers by gender



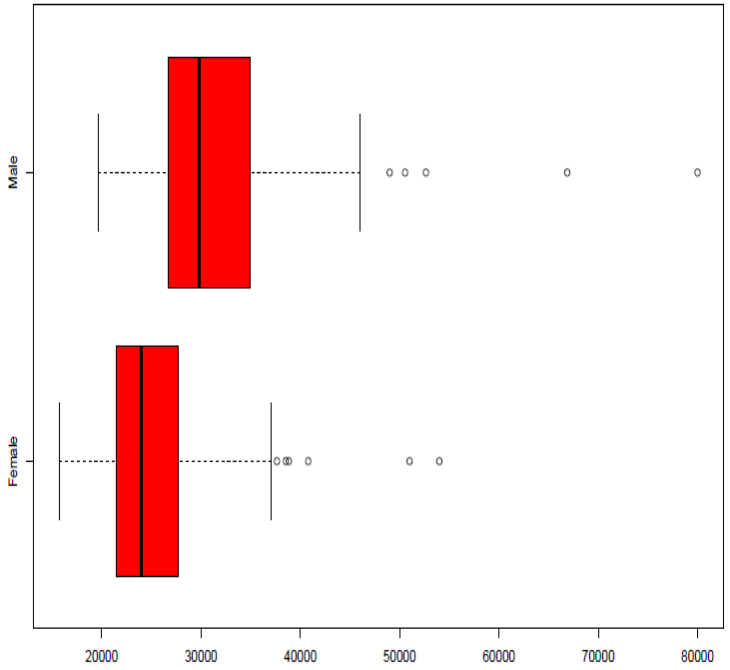
Education of managers by gender



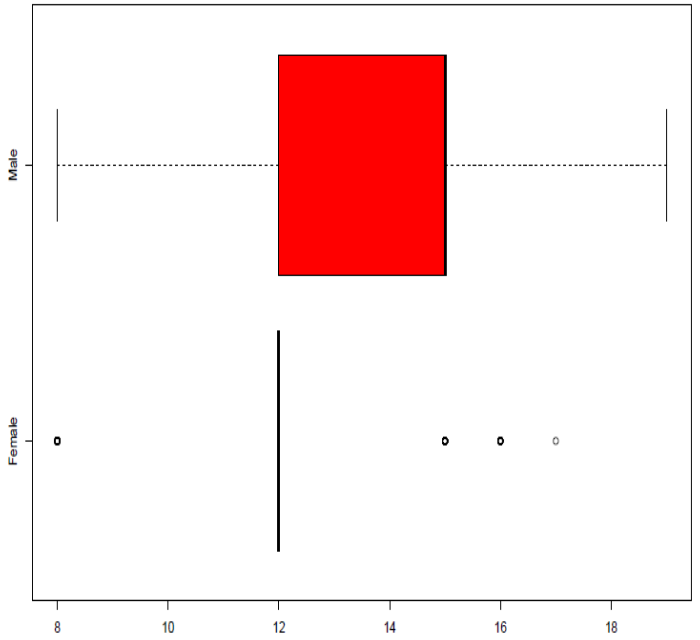
Male mangers are paid more but they are also spent more years at school compared to their female counter parts – However, there is still some level of discrimination against female mangers

Relationships between qualitative and quantitative variables

Current salary of clerical staff by gender



Education of clerical staff by gender



Male clericals are paid more but most of them are also spent more years at school compared to their female counter parts – However, there is still some level of discrimination against female mangers – more variability in salary

Relationships between qualitative variables

Distribution pattern for qualitative variable is given by the frequencies or relative frequencies of the observations for each of the categories

Job Category	Frequency	Relative frequency
Clerical	363	0.766
Custodial	27	0.057
Manager	84	0.177
Total	474	1.000

Gender	Frequency	Relative frequency
Female	216	0.456
Male	258	0.544
Total	474	1.000

Sometimes we are interested in looking the distributions of these two variables together – ***Joint distribution***

- Is there association between Job category and gender?
- Are males and females equally distributed in the different job categories?

Relationships between qualitative variables

- Joint distribution of two qualitative variables is the frequency of observations for two variables considered together as a combinations.
- Contingency table or cross-tabulation or two-way tables

Frequency

	Gender	
Job Category	Female	Male
Clerical	206	156
Custodial	0	27
Manager	10	74

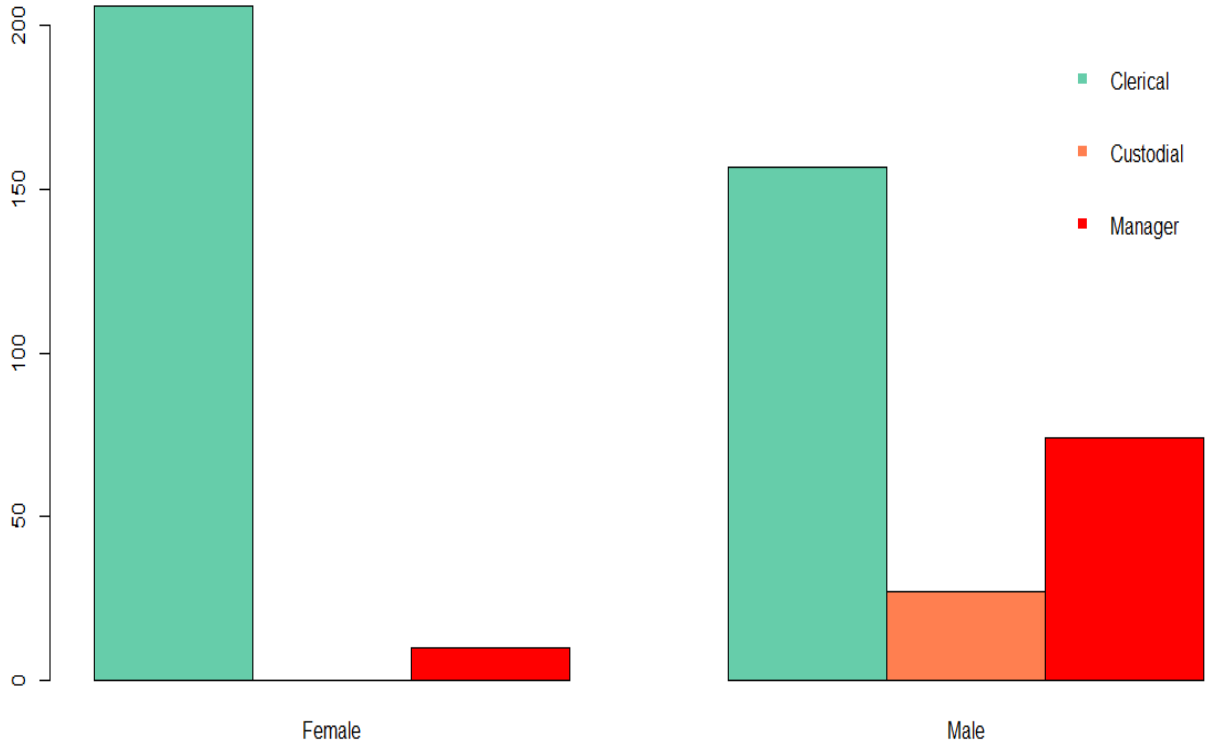
Relative frequency

	Gender	
Job Category	Female	Male
Clerical	0.435	0.331
Custodial	0.000	0.057
Manager	0.021	0.156

206/474

Relationships between qualitative variables

- Joint distribution of Job category and Gender



Relationships between qualitative variables

- Marginal distribution
- Look at one variable at a time

Job Category	Female	Male	Total
Clerical	206	156	362
Custodial	0	27	27
Manger	10	74	84
Total	216	257	473

Marginal distribution for job category

Marginal distribution for gender

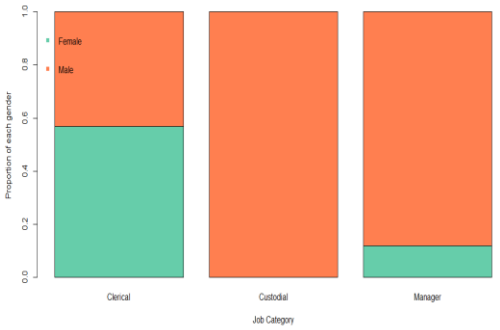
Relationships between qualitative variables

Conditional distribution

- Look at distribution one variable for each level of the other variable (conditioned)

	Gender	
Job Category	Female	Male
Clerical	206/362	156/362
Custodial	0/27	27/27
Manger	10/84	74/84

Are jobs distributed equally between male and female?



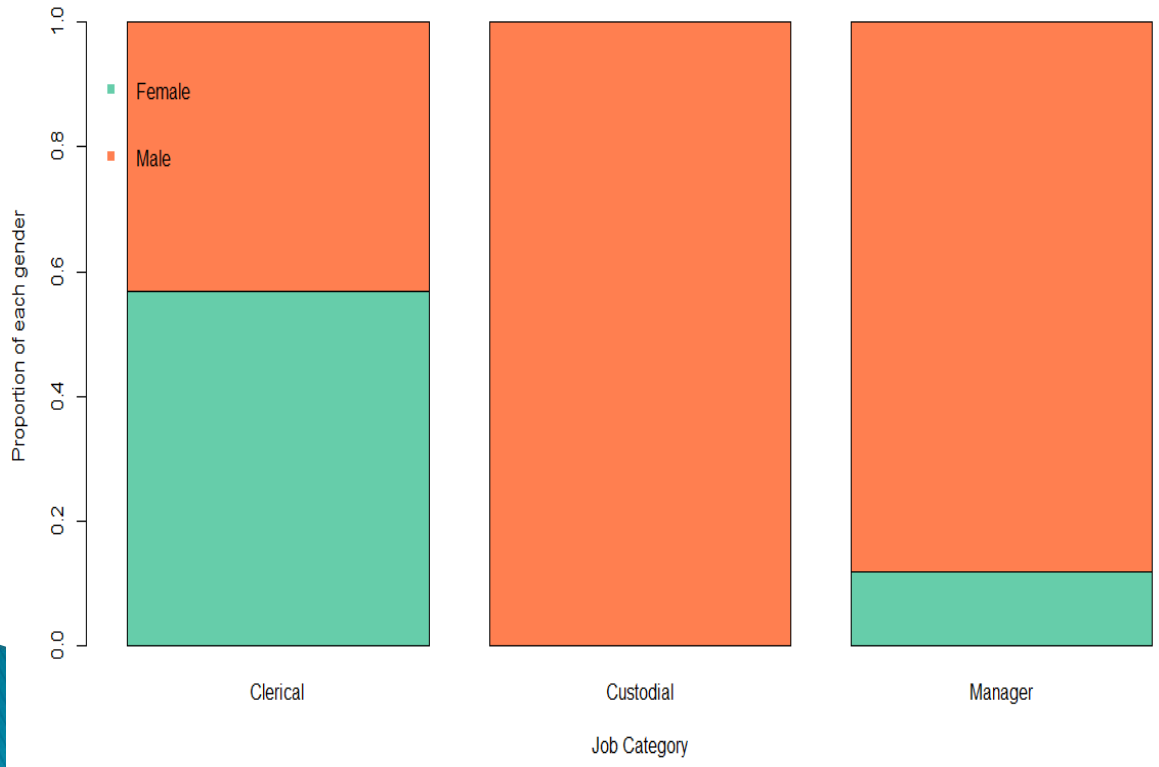
- All custodial jobs taken by men
- Men took 88% of managerial positions
- Women took about 57% of clerical jobs

Job Category	Female	Male
Clerical	56.9%	43.1%
Custodial	0.0%	100%
Manger	11.9%	88.1%

Relationships between qualitative variables

Conditional distribution

- Look at distribution one variable for each level of the other variable (conditioned)

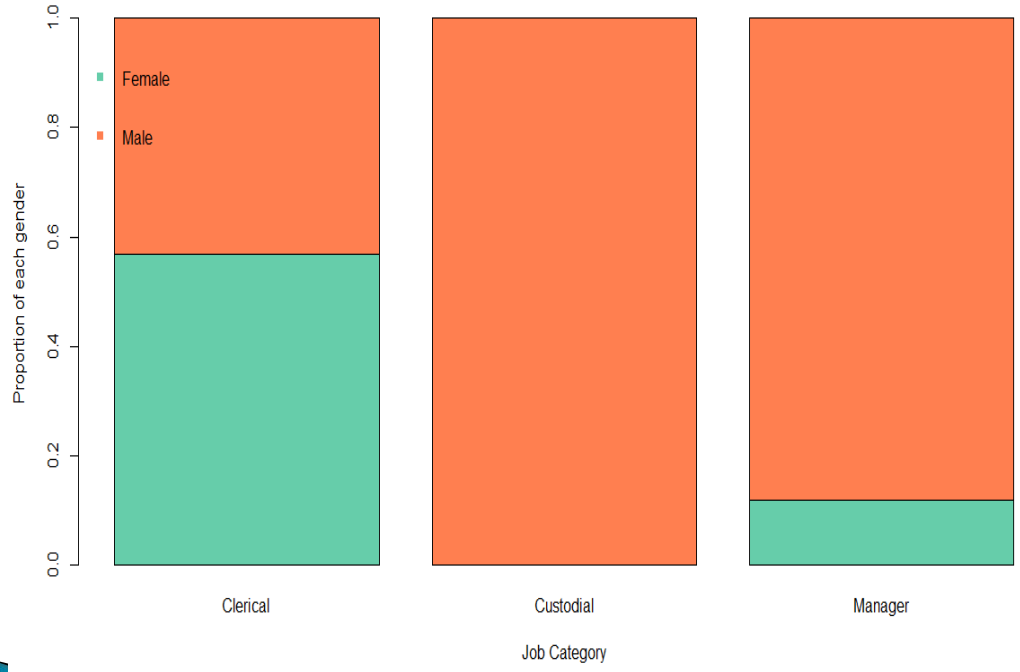


- All custodial jobs taken by men
- Men took 88% of managerial positions
- Women took about 57% of clerical jobs

Relationships between qualitative variables

Conditional distribution

- Look at distribution one variable for each level of the other variable (conditioned)



Two variables in a contingency table are **independent** if the conditional distribution of one variable is the same for the all values of the other variable – **Dependency/Association**

Distribution of the two genders in the three job categories differs – **Dependency/Association**

Relationships between quantitative variables

- **Quantitative variables**
  - Education level – number of years at school
  - Current salary (US$)
  - Beginning salary(US$)
  - Time on job (months)
  - Previous experience (years)
- Is there relationship between number of years at school and current salary?
- Does the current salary depend on beginning salary?
- Does the current salary depend on previous experience?
- Etc.

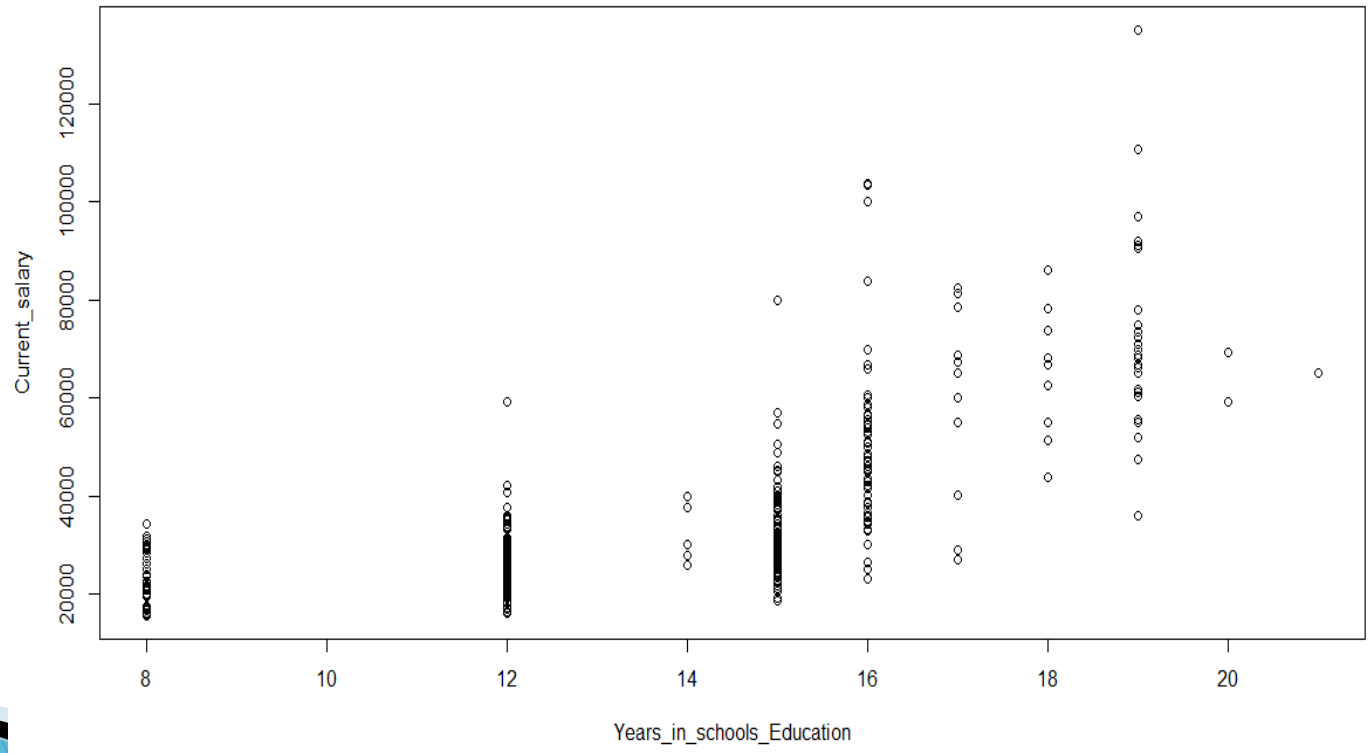
Scatterplots

- Basic graphical EDA technique – one variable on the x-axis, one on the y-axis and a point for each case in your dataset.
- Conventionally: X-axis = Explanatory/Independent/Input;
- Y-axis = Outcome/dependent variable.
- One or two additional categorical variables can be accommodated on the scatterplot by encoding the additional information in the symbol type and/or color.

Relationships between quantitative variables

Scatterplots

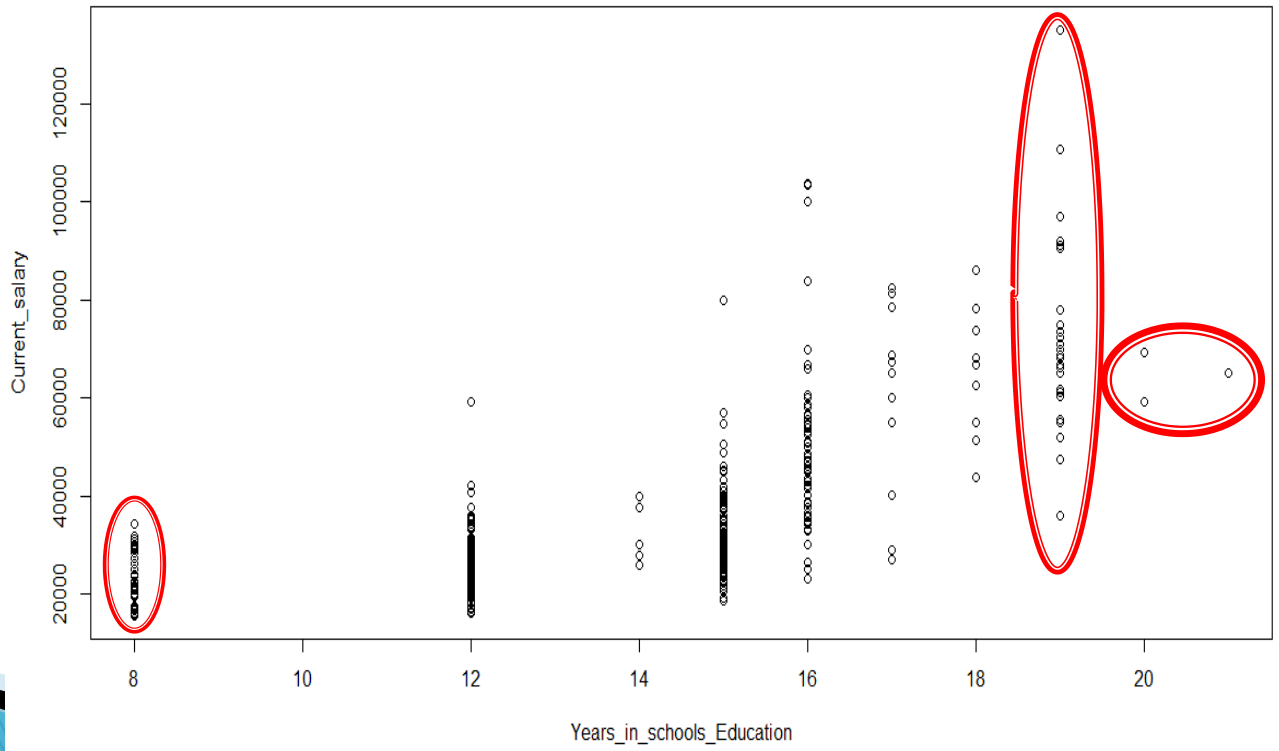
- Is there relationship between education level (years at school) and current salary?



Relationships between quantitative variables

Scatterplots

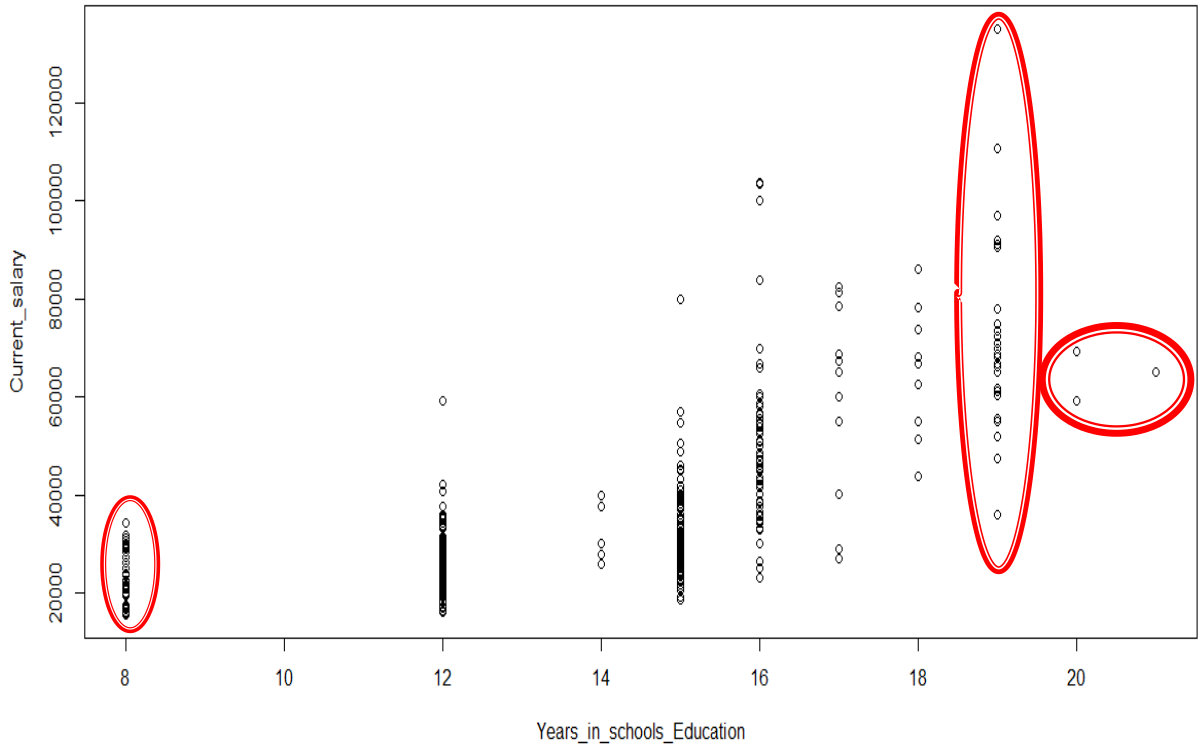
- Is there relationship between education level (years at school) and current salary?



Relationships between quantitative variables

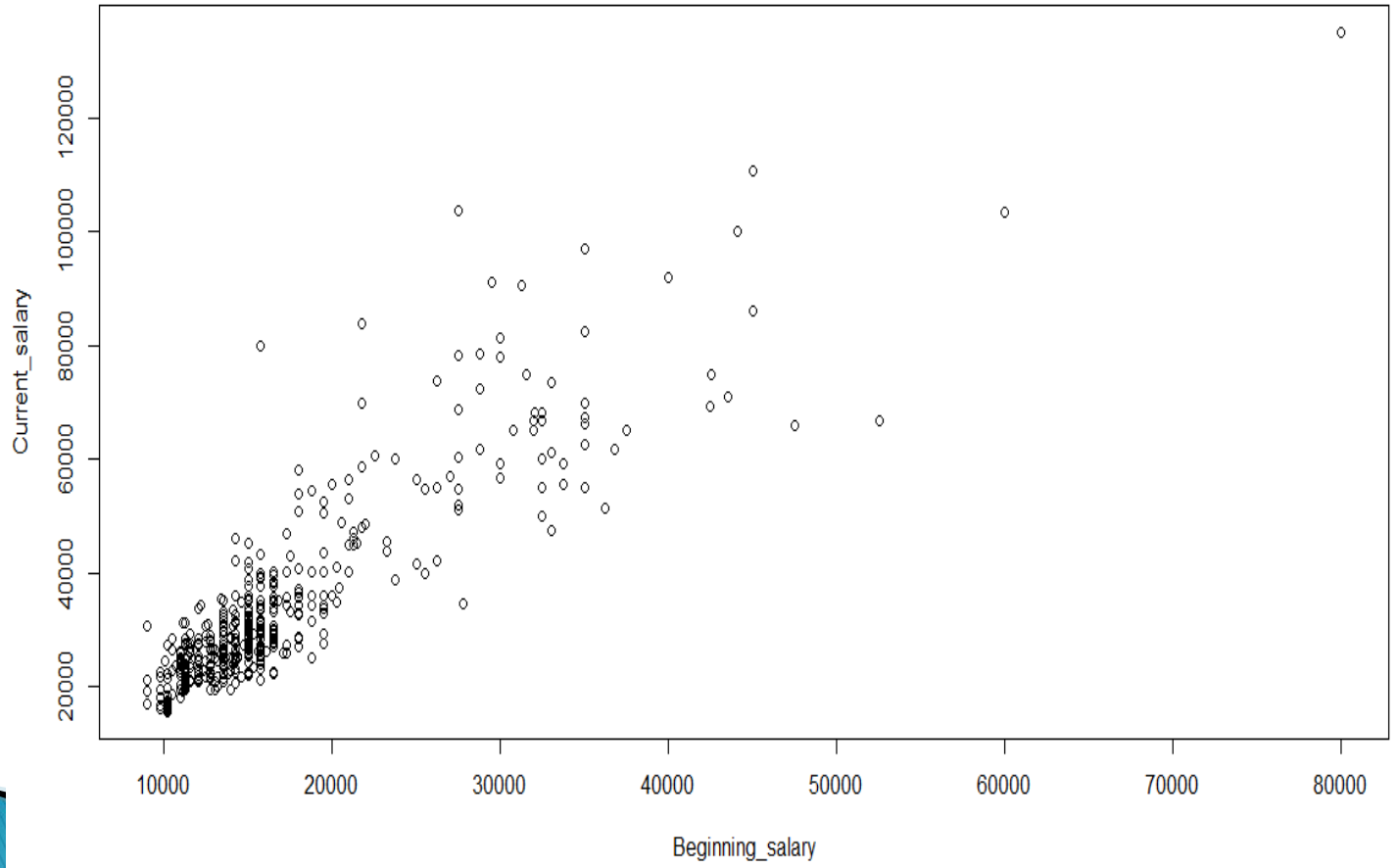
Years at school	8	12	14	15	16	17	18	19	20	21
Number of employee	53	190	6	116	59	11	9	27	2	1

Scatter plots



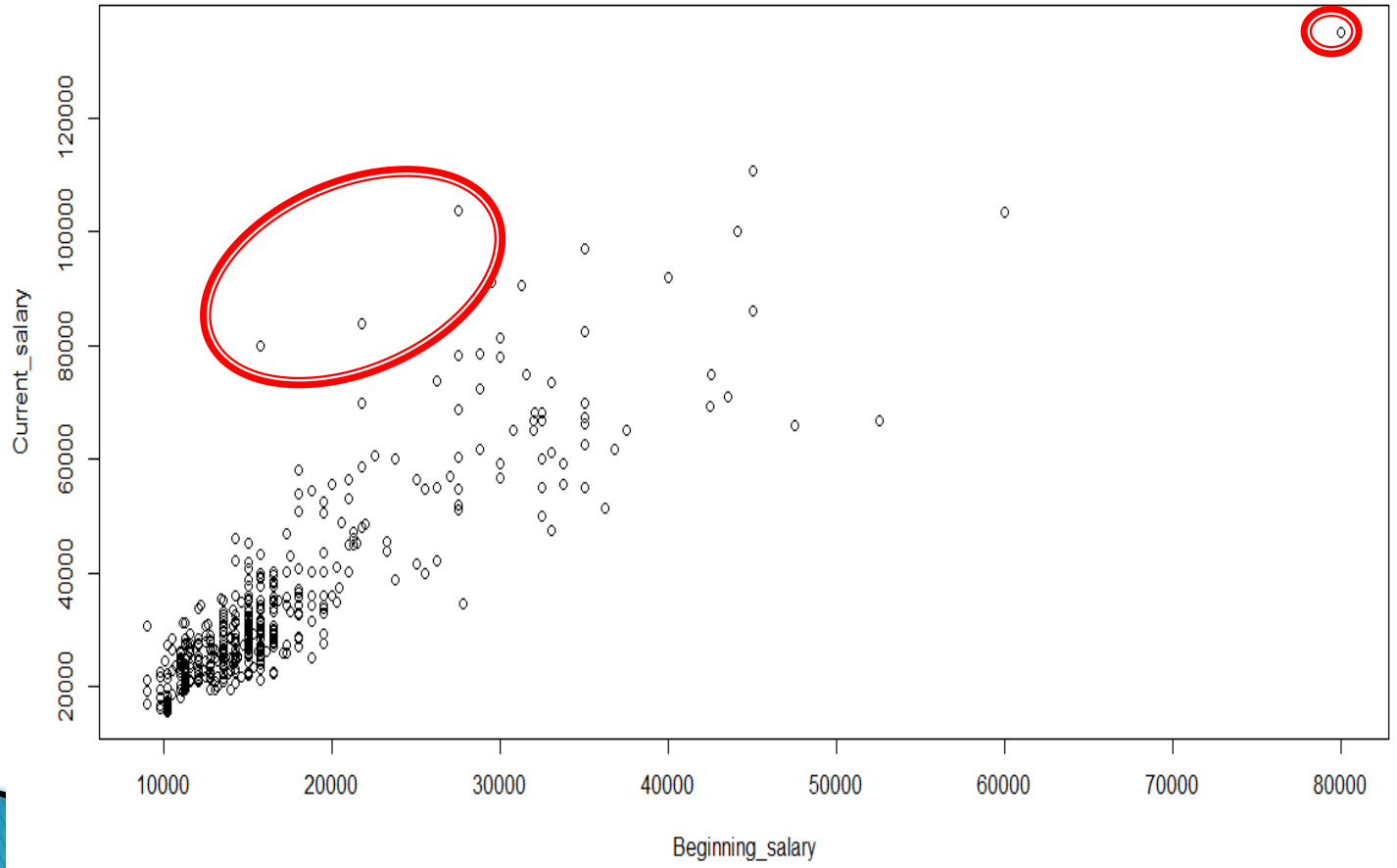
Relationships between quantitative variables

Scatterplots – Beginning salary versus current salary



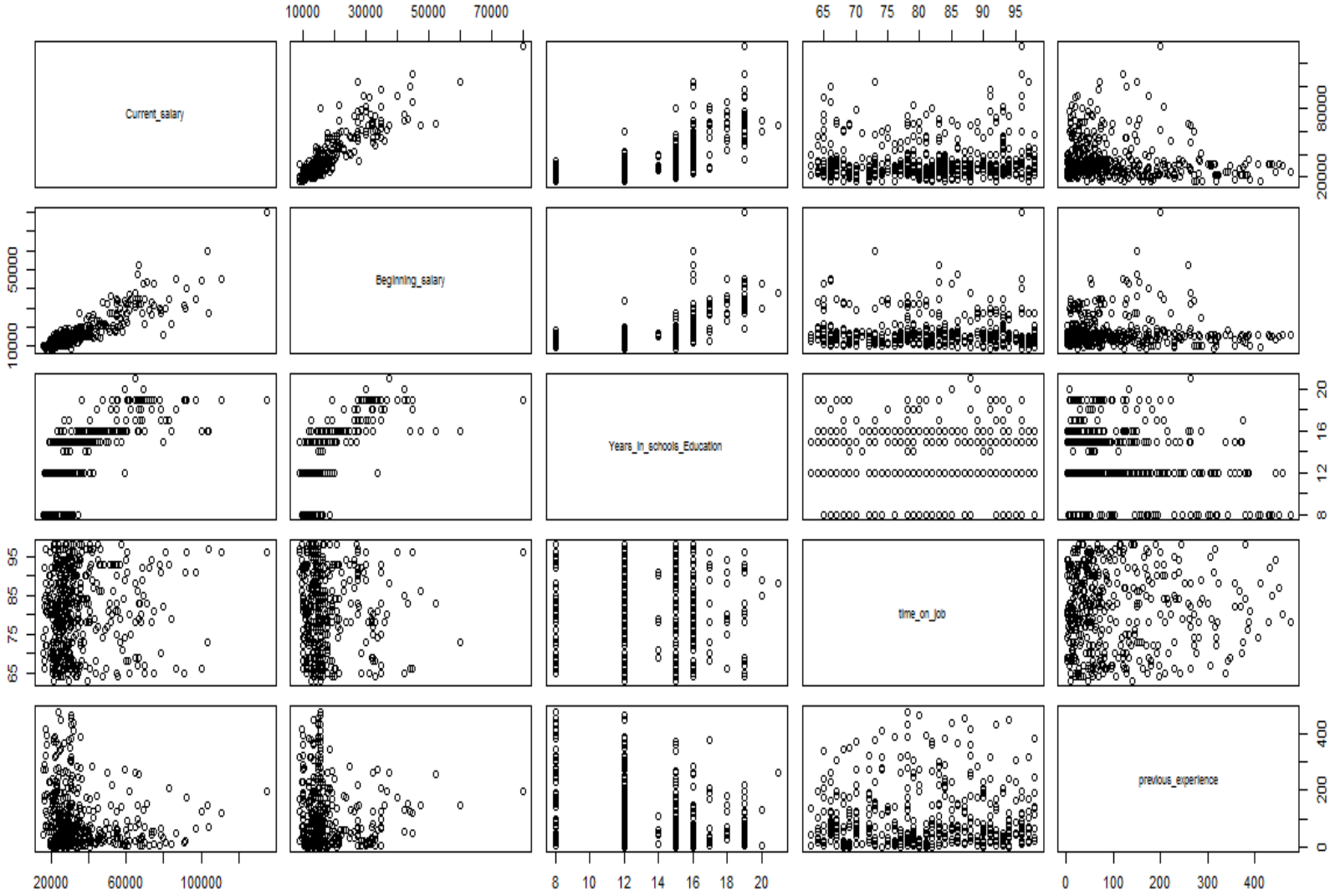
Relationships between quantitative variables

Scatterplots – Beginning salary versus current salary



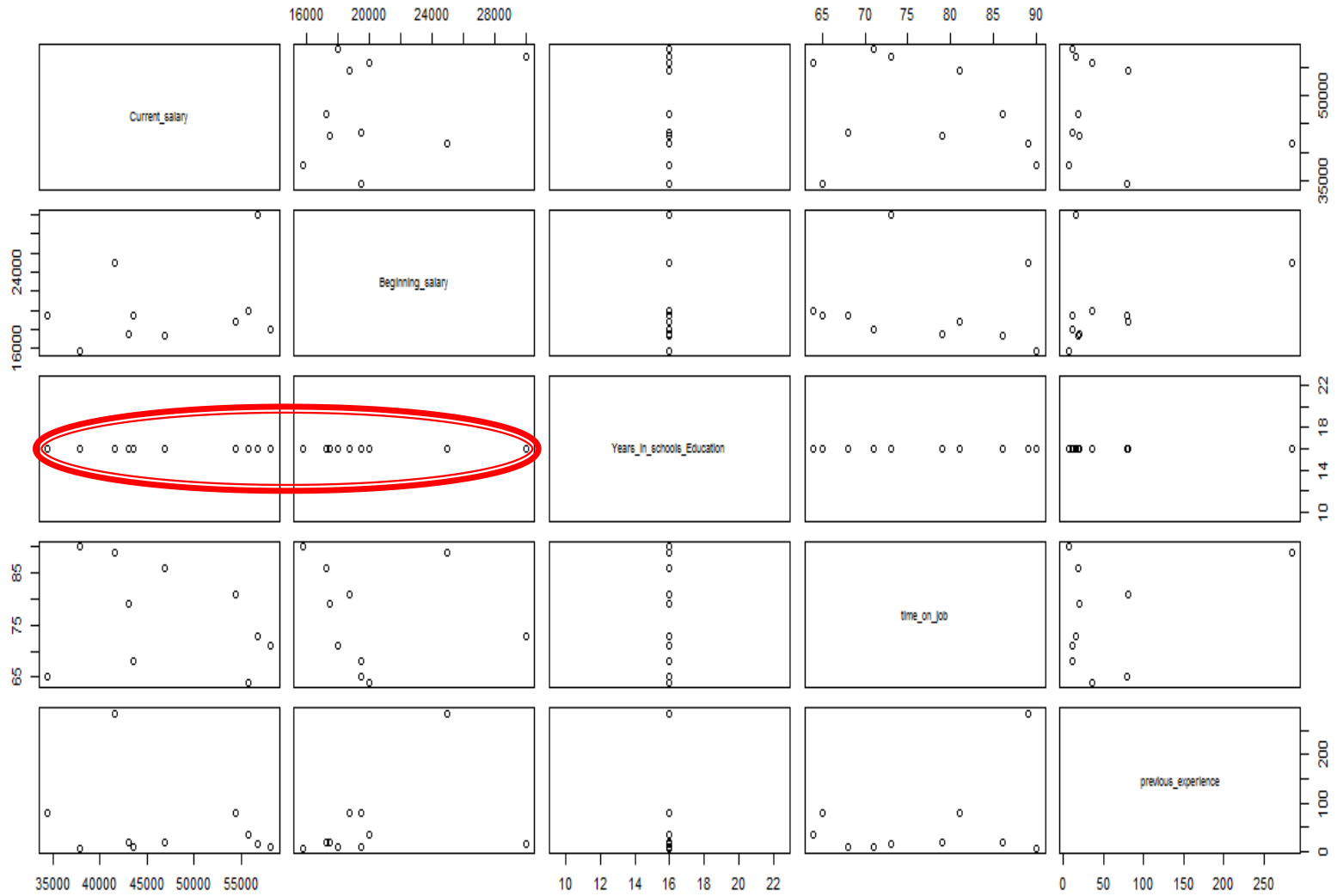
Relationships between quantitative variables

Multiple Scatterplots – all



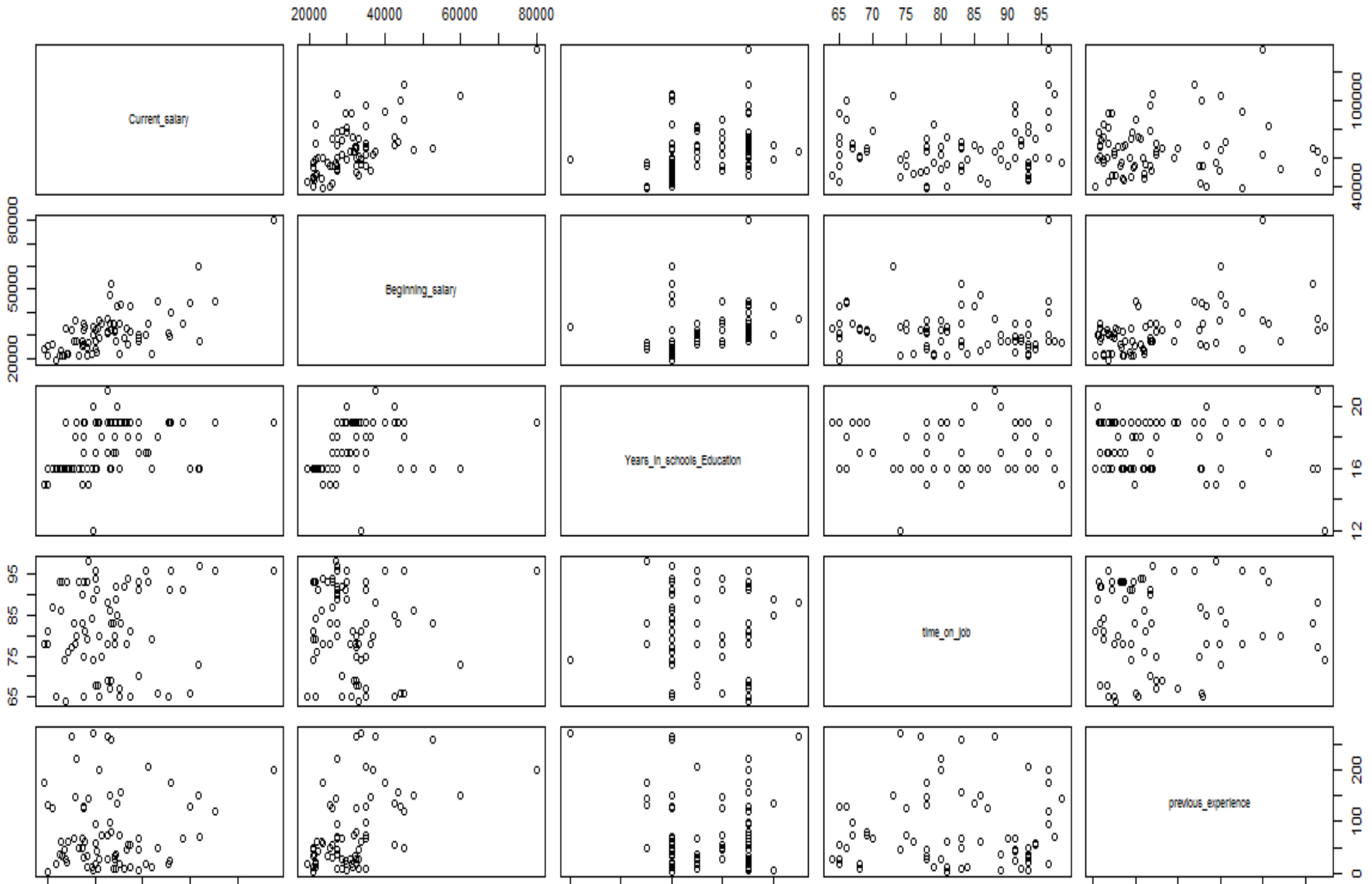
Relationships between quantitative variables

Multiple Scatterplots –



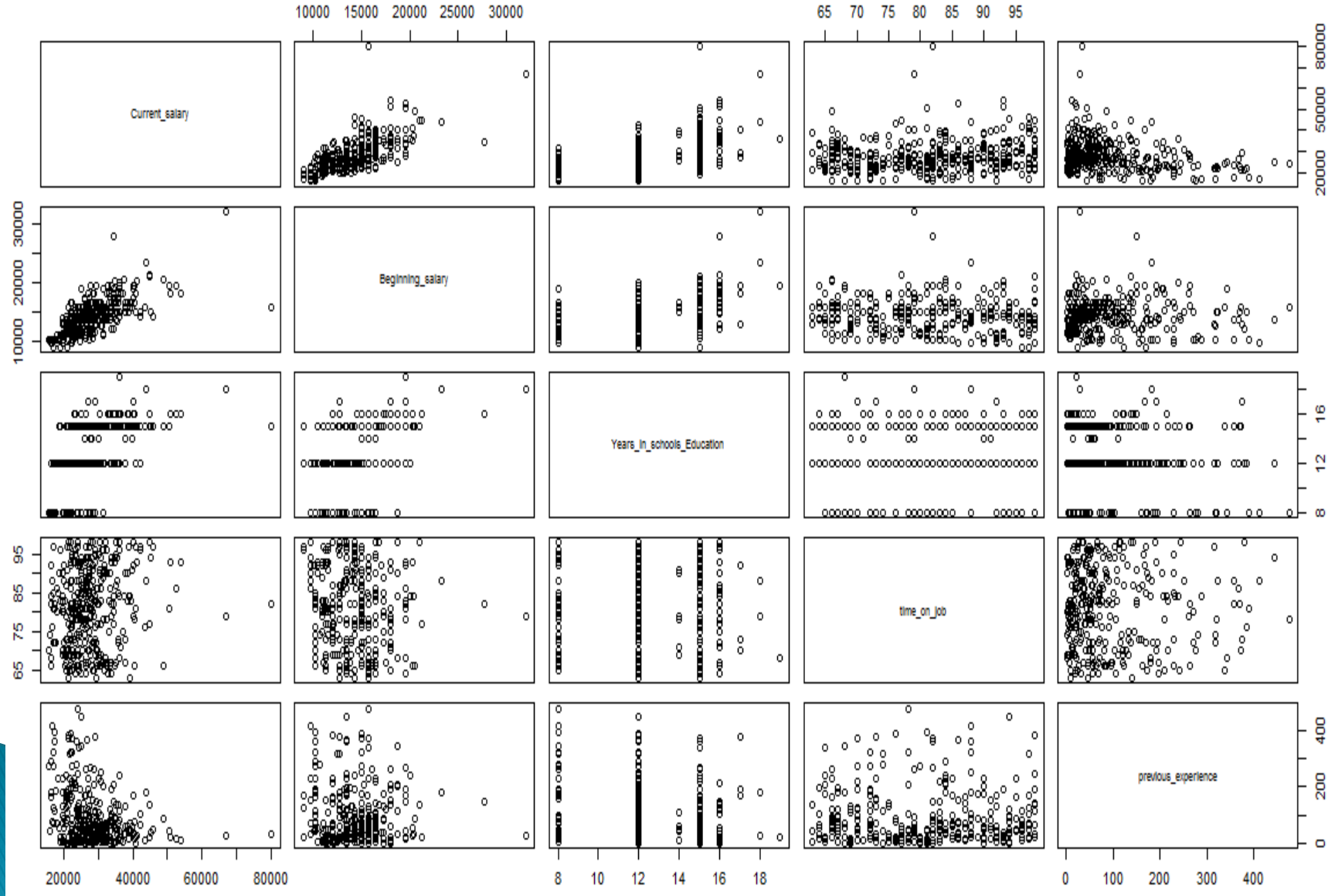
Relationships between quantitative variables

Multiple Scatterplots – Male



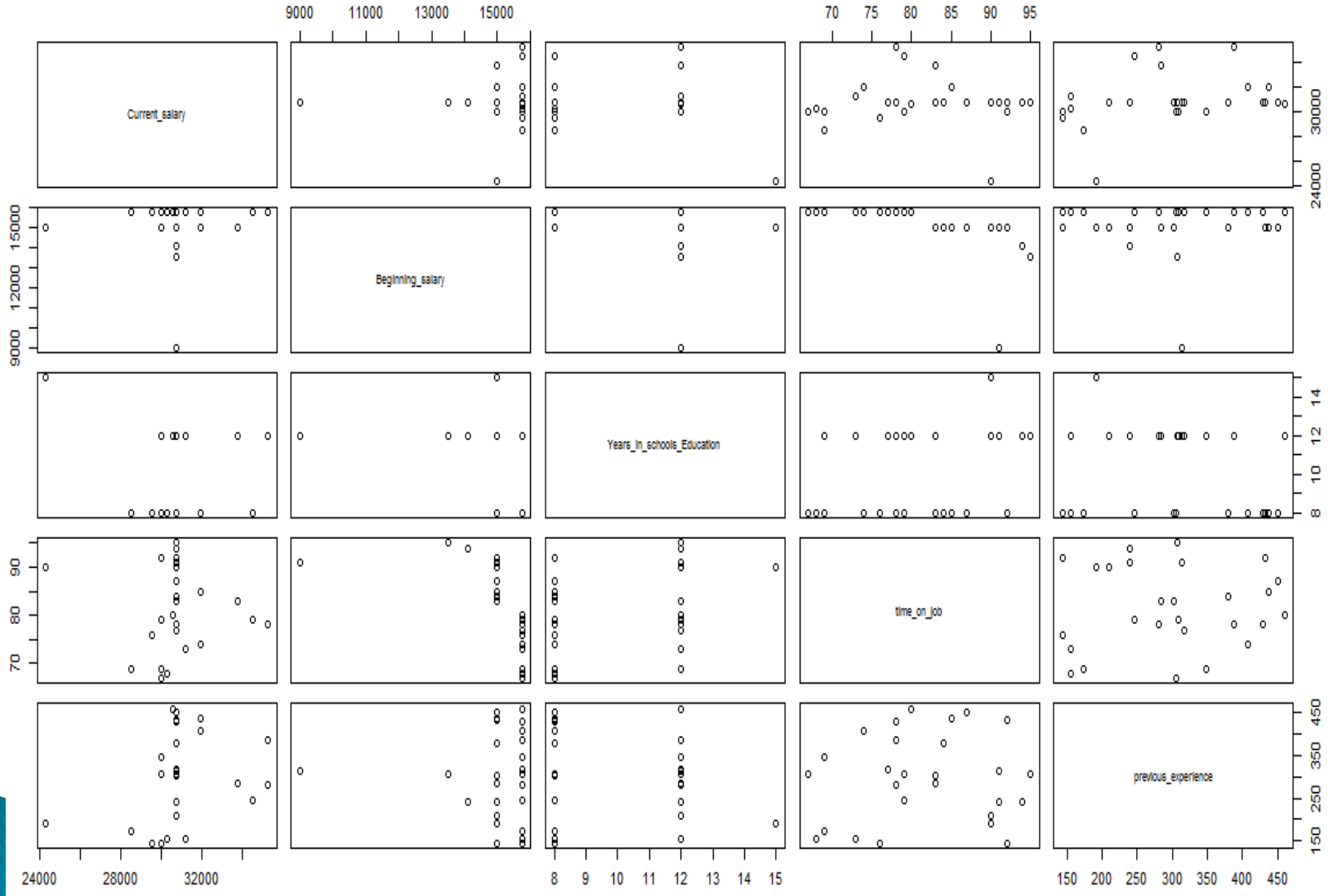
Relationships between quantitative variables

Multiple Scatterplots –



Relationships between quantitative variables

Multiple Scatterplots –



Relationships between quantitative variables

Correlation and covariance

- Covariance is a measure of how much two variables "co-vary", i.e., how much (and in what direction) should we expect one variable to change when the other changes.
- Unit of a covariance are the products of the units of the two variables. Because it unit dependent is hard to make comparison
- **Correlation is a standardized measure is which scale independent – unit less – good for comparison**

$$Cov(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

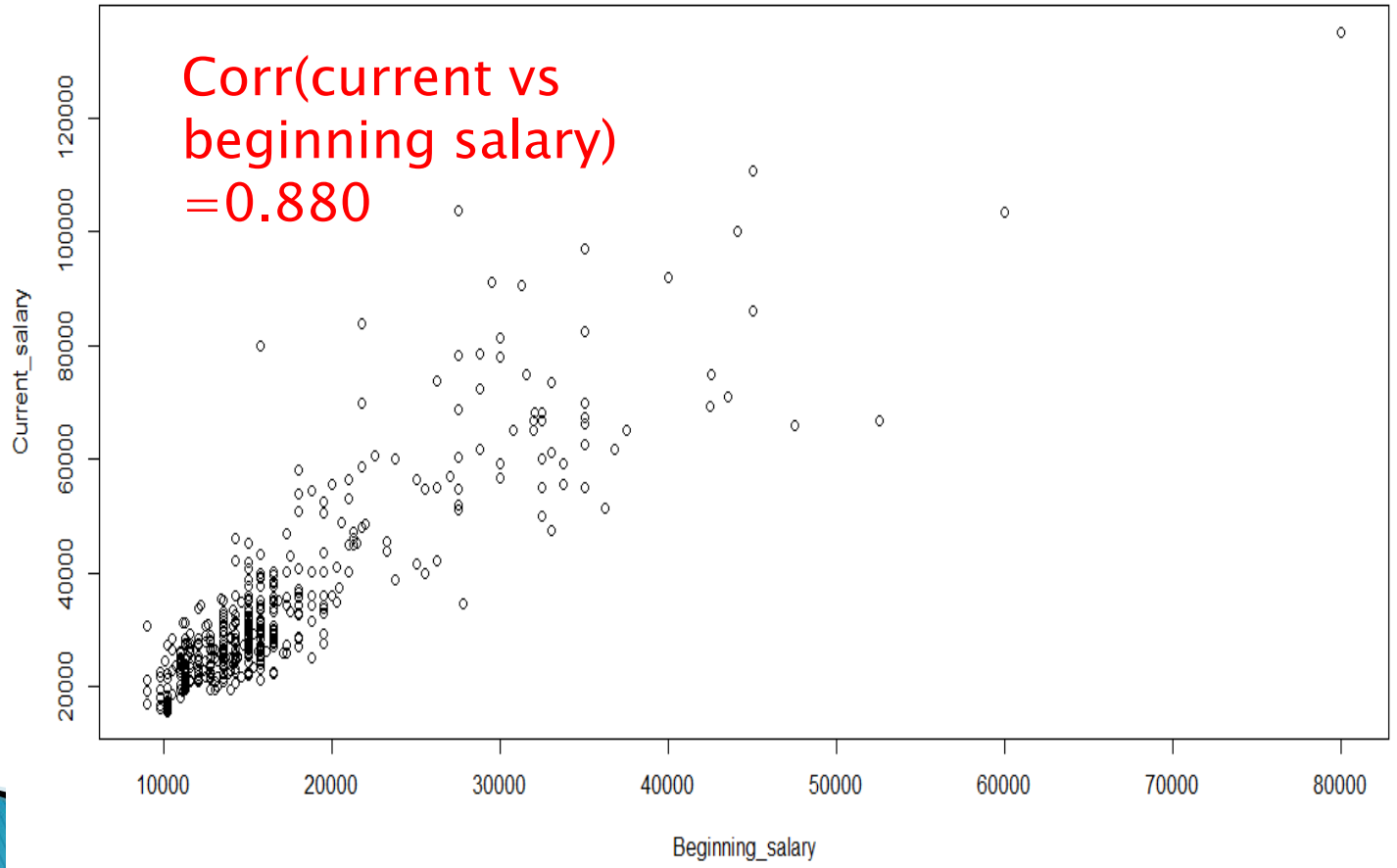
$$\text{Corr}(x,y) = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 \sum(y_i-\bar{y})^2}}$$

$$\text{Corr}(x,x) = \frac{\sum(x_i-\bar{x})(x_i-\bar{x})}{\sqrt{\sum(x_i-\bar{x})^2 \sum(x_i-\bar{x})^2}}$$

Correlation between a variable and itself is 1 – perfect correlation

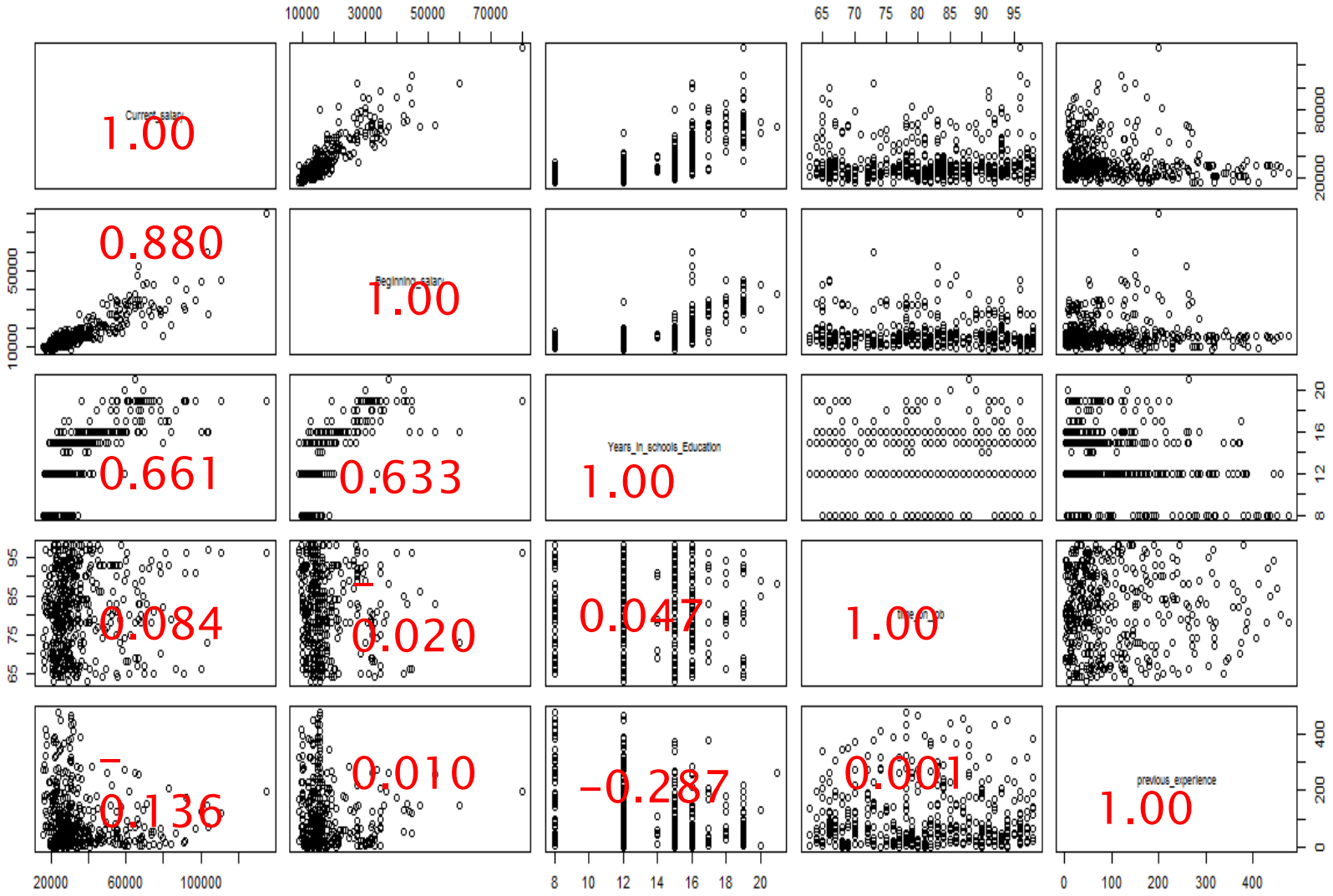
Relationships between quantitative variables

Correlation- Beginning salary versus current salary



Relationships between quantitative variables

Correlations- all observations



Relationships between quantitative variables

	Correlation between current salary and other variables for different data sets							
	Combined data	Manager	Clerical	Custodial	Female Manager	Male Manager	Female Clerical	Male Clerical
Beginning Salary	0.880	0.693	0.683	0.077	0.295	0.657	0.690	0.518
Education	0.661	0.397	0.517	-0.107	-	0.337	0.457	0.436
Time on Job	0.084	0.168	0.147	-0.096	-0.269	0.144	0.207	0.076
Previous experience	-0.136	0.094	-0.272	0.284	-0.258	0.085	-0.312	-0.320

THANK YOU FOR LISTENING

Let us turn to R