Data Analytics

Lecture No: 02

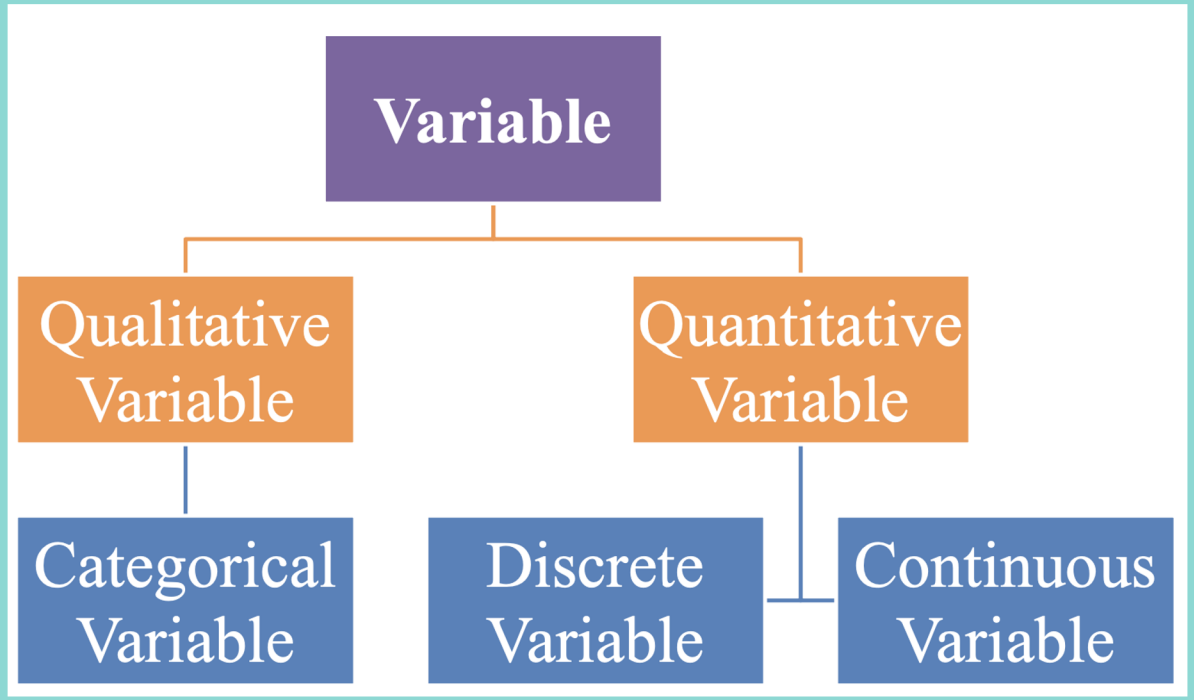
Variables

- Any characteristic of interest which takes on different values is called variable.

Or

- A variable is any characteristics, number, or quantity that can be measured or counted.
- A variable may also be called a data item.
- For example, price of a commodity at different places in Peshawar city, profit of a business firm at different months of a year, production, cost, temperature, sale of a market, consumption etc.
- Variable is broadly divided into qualitative and quantitative variables.

Types of Variable



Cont'd

Qualitative data

- Qualitative data are generally described by words or letters.
- They are not as widely used as quantitative data because many numerical techniques do not apply to the qualitative data.
- For example, it does not make sense to find an average hair color or blood type. Satisfaction level from teacher. Intelligence, Beauty, and Gender.

Qualitative data can be separated into two subgroups:

- **Dichotomic:** (if it takes the form of a word with two options (gender - male or female)
- **Polynomic:** (if it takes the form of a word with more than two options (education - primary school, secondary school and university).

Types of variable

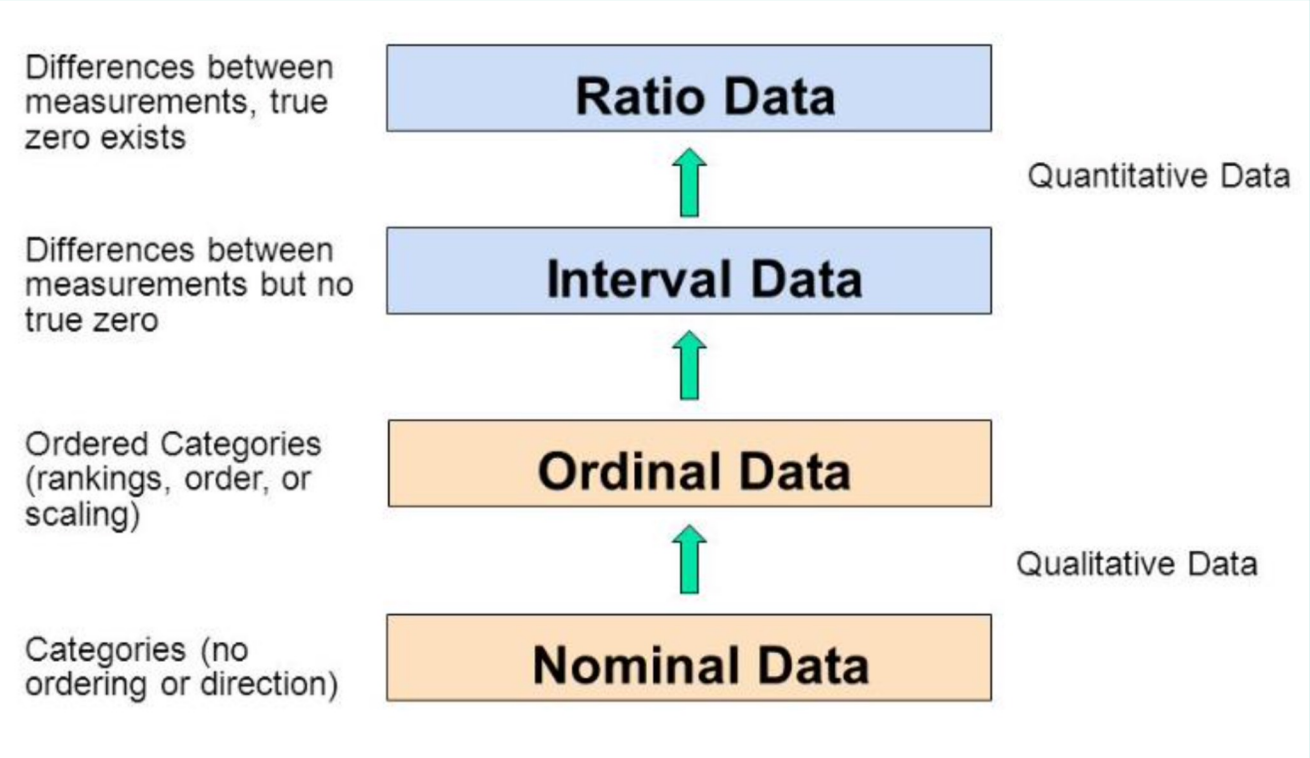
Quantitative data

- Quantitative data are always numbers and are the **result of counting or measuring** attributes of a population.
- Numerical data

Quantitative data can be separated into two subgroups:

- **discrete:** (if it is the result of counting (the number of students of a given ethnic group in a class, the number of books on a shelf, ...)
- **continuous:** (if it is the result of measuring (distance traveled, amount of milk, Temperature weight of luggage, ...)

Cont'd


Differences between measurements, true zero exists
Ratio Data
Quantitative Data
Differences between measurements but no true zero
Interval Data
Ordered Categories (rankings, order, or scaling)
Ordinal Data
Qualitative Data
Categories (no ordering or direction)
Nominal Data

Qualitative Variable

- **Nominal:**
  - Define categories **without a natural order or rank**
  - Frequency and Proportions
  - Pie chart, line chart and bar chart
  - Examples of nominal variable *blood type*, *gender*, *eye color*, *political party*

- **Ordinal:**
  - The variables in ordinal data are **listed in an ordered manner.**
  - The ordinal variables are usually numbered, so as to indicate the order of the list. (1 to 5 scale)
  - We cannot say anything about the intervals between the rankings.
  - Frequency and Proportions
  - Ordinal data cannot be represented in Pie chart
  - Examples of ordinal variables: socio economic status (“low income”,”middle income”,”high income”), education level (“high school”,”BS”,”MS”,”PhD”),

Cont'd

- **Interval/ Ratio:**
  - Most precise measurement
  - Scale, Qualitative and Parametric
  - Discrete and Continuous data
  - Mean, median Standard deviation
  - Line and bar chart
- **For example** very hot, hot, cold, very cold, warm are all nominal data when considered individually. But when placed on a scale and arranged in a given order (very hot, hot, warm, cold, very cold), they are regarded as ordinal data.

Invoice Data:

Invoice No	Customer Name	Product	Qty	Price	Customer Satisfaction
101	Mr. A	Laptop	1	70,000	5
102	Mr. B	Tablets	1	40000	4

Types of Variable

OK to compute	Nominal	Ordinal	Interval	Ratio
Frequency Distribution	Yes	Yes	Yes	Yes
Median and Percentiles	No	Yes	Yes	Yes
Add or Subtract	No	No	Yes	Yes
Mean, Standard deviation, standard error of the mean	No	No	Yes	Yes
Ratio, Coefficient of variation	No	No	No	Yes

Cont'd

Dependent and Independent Variables

- A type of variable which is **influenced** by other variable/variables is called dependent variable. It is also called random or stochastic variable.

OR

- A variable which depends on one or more other variables is called dependent variable.

OR

- A variable of primary interest that lends itself for investigation as a function of other cause variables is known as dependent variable.

Example

- In economics, consumption of a commodity (say apple) depends upon the income, household size, and price etc of the commodity. In this example, consumption of apple is a dependent variable which will vary from one family to other family; while the other variables like income, household size and price are independent variables.

- A variable which influence a dependent variable in either direction (positive or negative) is called ***independent variable***.

Data

- **Grouped Data:**
  - Grouped data means the data (or information) given in the form of class intervals such as 0-20, 20-40 and so on.
  - Grouped data is data that has been bundled together in categories.
  - Histograms and frequency tables can be used to show this type of data.

- **Ungrouped Data:**
  - Ungrouped Data is defined as the data given as individual points (i.e. values or numbers) such as 15, 63, 34, 20, 25, and so on.
  - Ungrouped data is the data you first gather from an experiment or study.
  - The data is raw — that is, it's not sorted into categories, classified, or otherwise grouped.
  - An ungrouped set of data is basically a list of numbers.

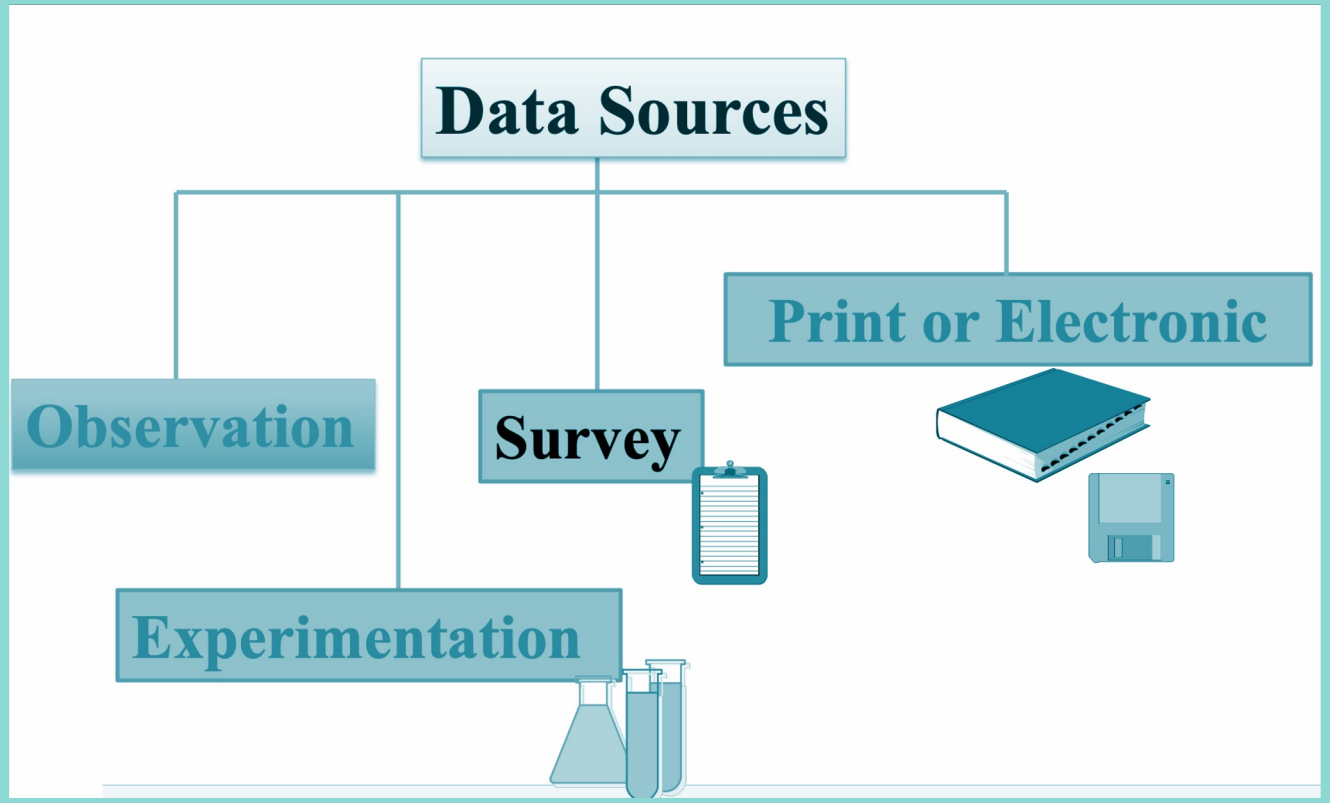
Cont'd

- Grouping of data plays a significant role when we have to deal with large data. This information can also be displayed using a pictograph or a bar graph.
- Data formed by arranging individual observations of a variable into groups.
- A frequency distribution table of these groups provides a convenient way of summarizing or analyzing the data is termed as grouped data.
- Suppose we have a data ranges from 0 to 50 like 2, 17, 0, 1, 8, 19, 43, 2, 1, 32, and so on. In this case, we can group the data into classes such as 0-10, 10-20,...,40-50. This is a simple example of grouped data.

The main advantages of grouping data are:

- Assist us in concentrating on essential subgroups mainly and overlooks trivial ones.
- Helps in increasing the efficiency and correctness of the required estimation

Sources of Data



Types of Data Collection

- **Primary Data**
  - The data which is collected for the first time from its source, is called primary data.

- **Secondary Data**
  - When the primary data is passed through any sort of statistical or mathematical treatment, the data is known as secondary data.

OR

  - The data that are collected and compiled by an outside source or by someone in the organization who may later provide access to the data to other users.

Data

Time Series Data:

- The data collected at different interval of time regarding a commodity or group of commodities (or organization/firm) is called time series data. For example,

Time Series data of a company showing profit, production and sale.

Year	Profit	Production	Sale
1990	12	120	110
1991	13	140	132
1992	14	150	145
1993	13.5	140	123
1994	10	103	90
1995	11	115	100
1996	12.5	123	122
1997	13.8	140	135
1998	15	160	145
2000	11.6	120	115
2001	15	162	150
2002	16	165	145

Data

Cross Sectional Data

Widely dispersed data (such as) relating to one period, or data related to households, data collected from the field survey i.e. monthly profit of the selected stores, or monthly profit of different companies related to only one period etc

Cross sectional data of 12 different households showing profit, production and sale.

Household	Profit	Production	Sale
1	12	120	110
2	13	140	132
3	14	150	145
4	13.5	140	123
5	10	103	90
6	11	115	100
7	12.5	123	122
8	13.8	140	135
9	15	160	145
10	11.6	120	115
11	15	162	150
12	16	165	145

Data

Panel Data

Panel data is data from a (usually small) number of observations over time on a (usually large) number of cross-sectional units like individuals, households, firms, or governments.

Panel data of three different firms showing, production and sale during 2000-2002.

Firm	Year	Production	Sale
1	2000	120	110
1	2001	140	132
1	2002	150	145
2	2000	140	123
2	2001	103	90
2	2002	115	100
3	2000	123	122
3	2001	140	135
3	2002	160	145
4	2000	120	115
4	2001	162	150
4	2002	165	145

Data Collection Tools

- **Primary Data**
  - Direct personal investigation (Schedule) Questionnaire methods
  - Through Enumerators etc.

- **Secondary Data**
  - Printed materials
  - Bureau of Statistics
  - State Bank of Pakistan Commercial and Research Journals

Frequency

- Repetition of an observation in a data set is called frequency of that particular observation/data point/individual.

OR

- Total number of observations in a class is called the frequency of that class.
- For example, consider the following data showing the monthly salaries of 50 employees of a certain University. In this example, 20 is the frequency of the class (employees) having salary Rs. 40, 000 per month, and 3 is the frequency of the employees drawing Rs. 90,000 per month salary.

Salary(K)	40	50	60	70	80	90
No of Employees	20	10	8	5	4	3

Data Representation

- Diagrams and Graphs.
- Simple Bar Diagram.
- Multiple Bar Diagram.
- Component Bar Diagram.
- Pie Chart.
- Frequency Curves.
- Graphical Display of Data

Simple Bar Diagram

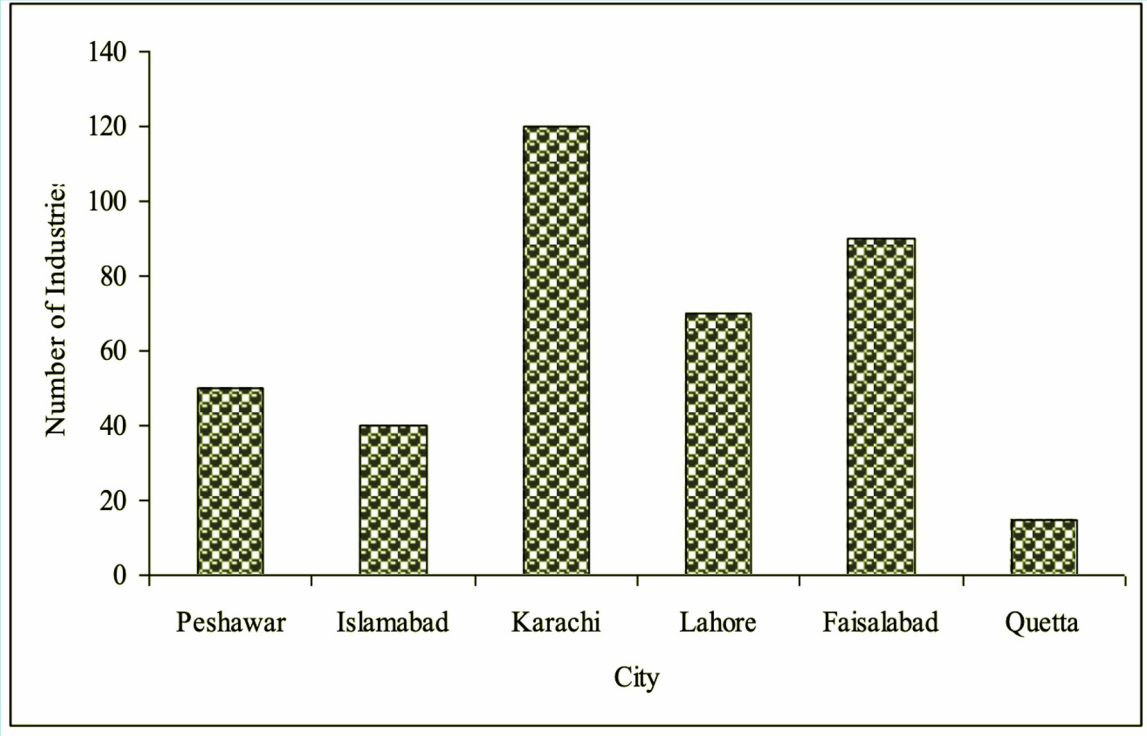


Figure: Summary of the number of industries in different cities of Pakistan

Multiple Bar Diagram

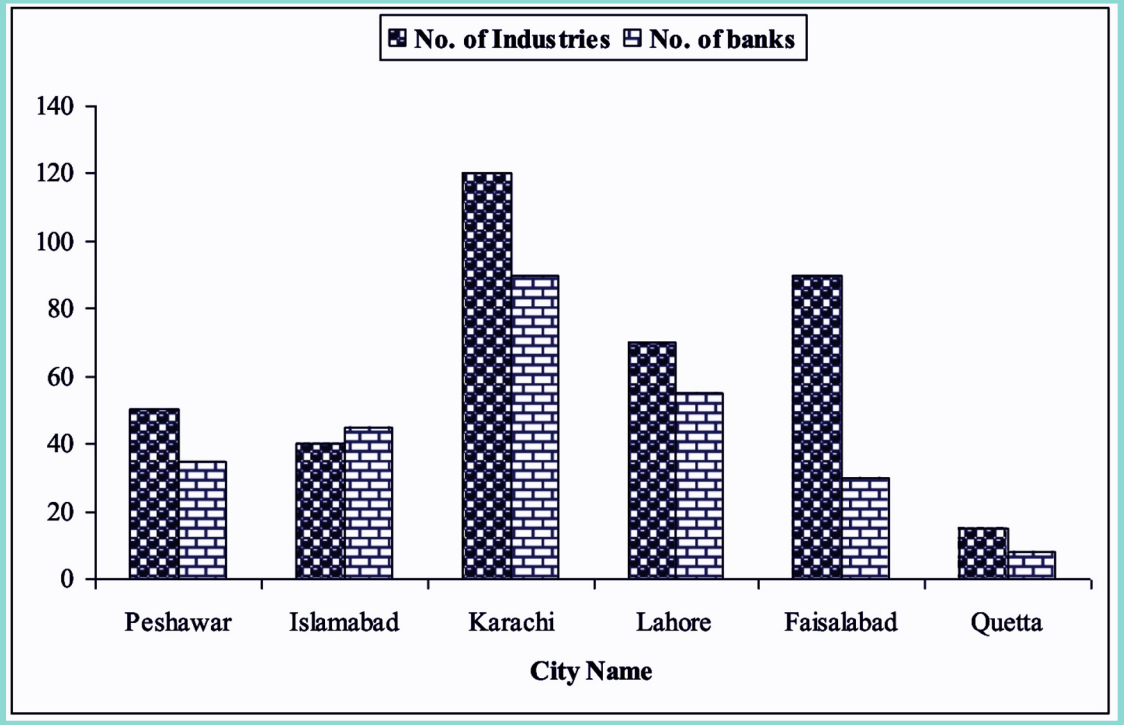


Figure: Summary of the number of industries and Banks in different cities of Pakistan

Component Bar Diagram

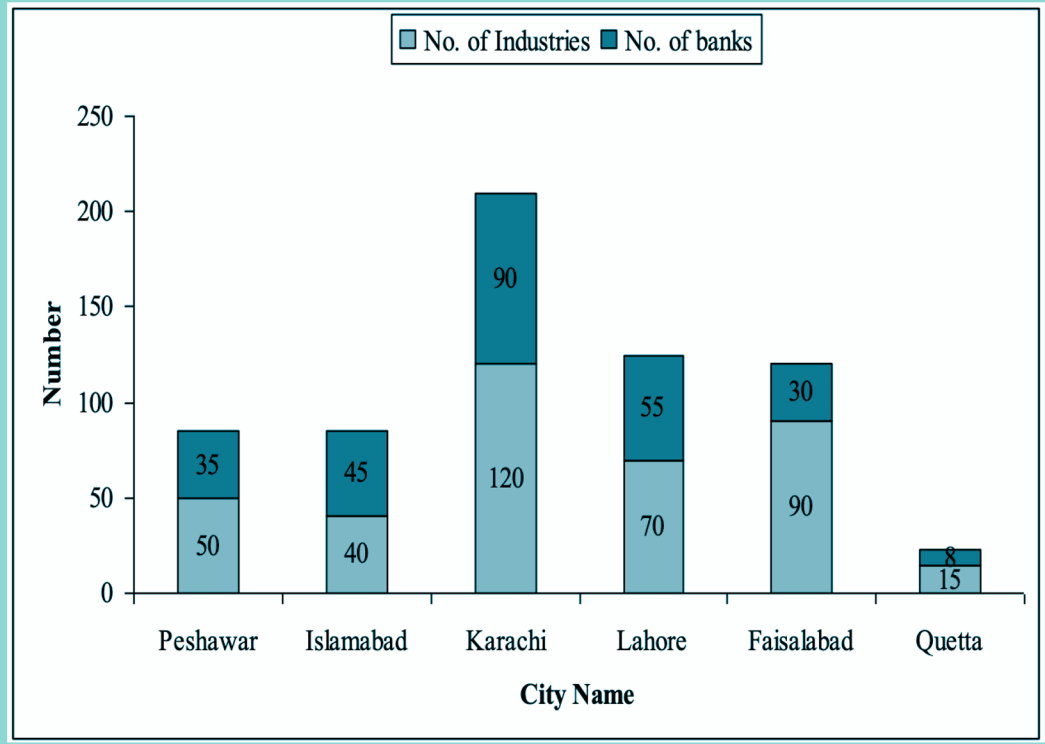


Figure: Summary of the number of industries and Banks in different cities of Pakistan

Pie Chart

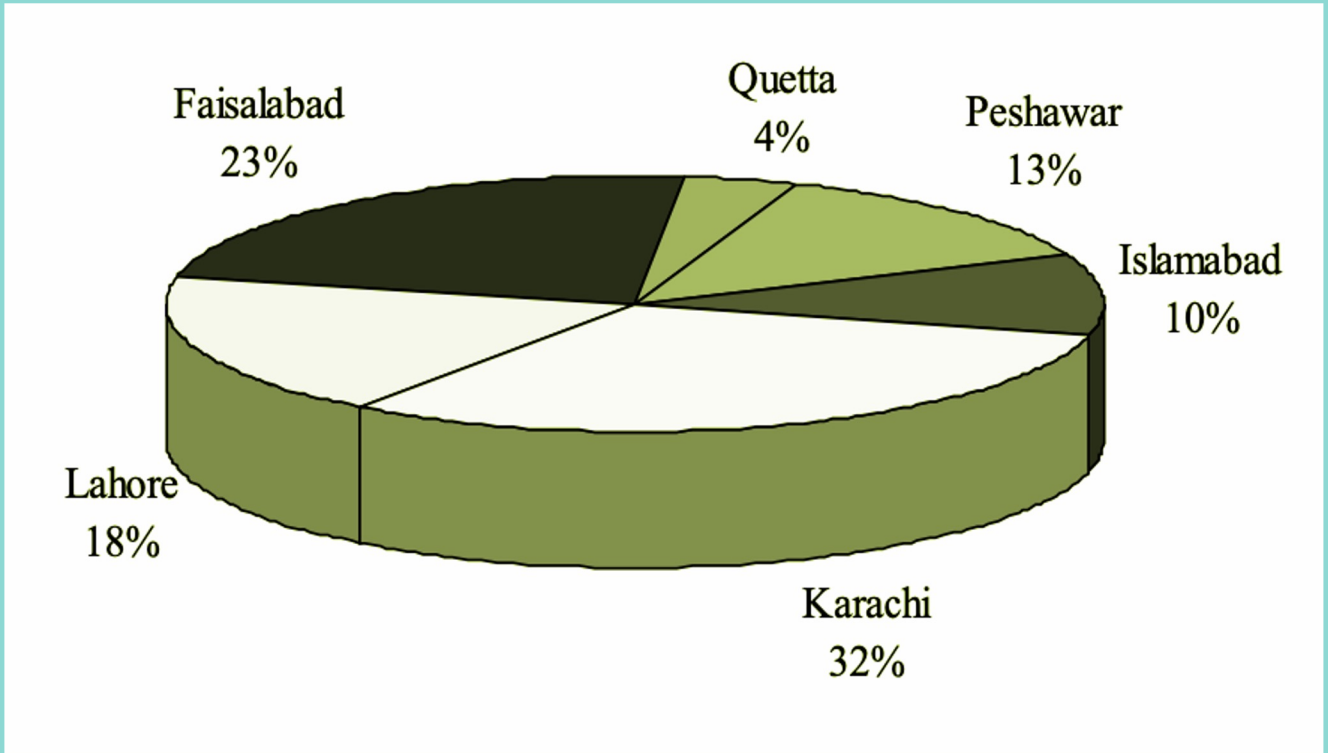
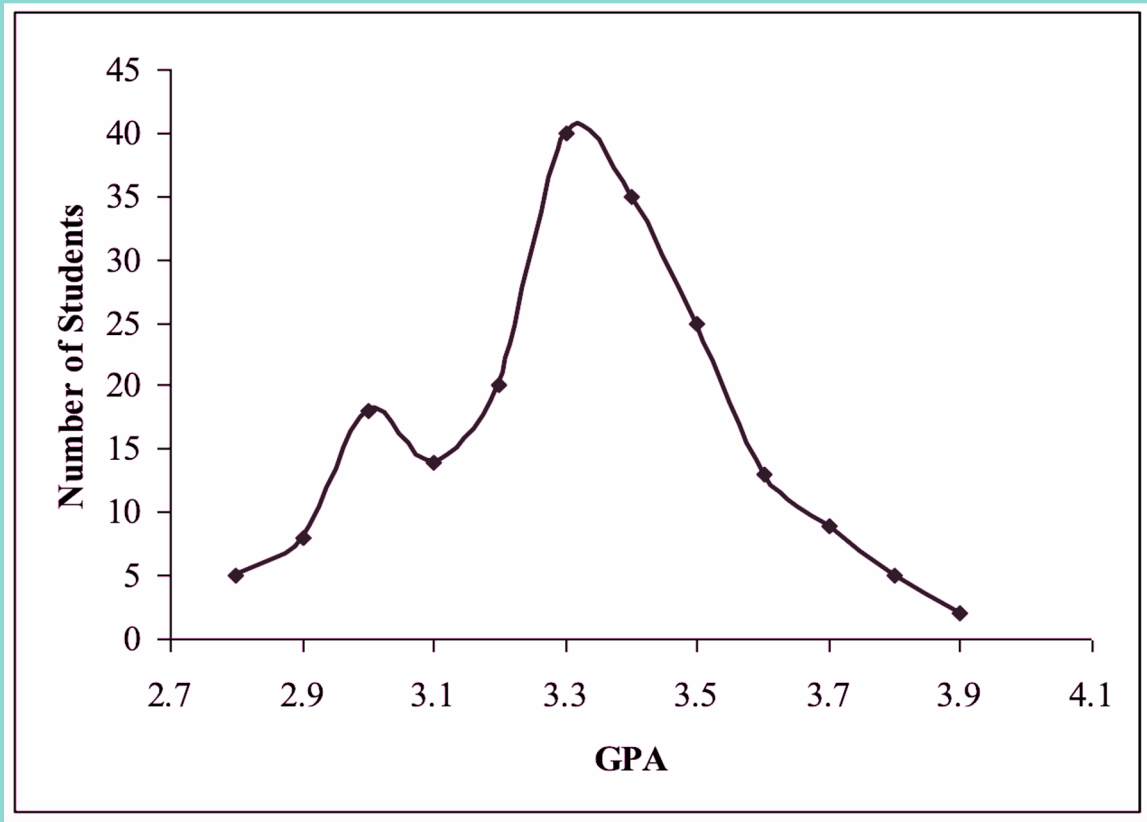


Figure: Summary of the number of industries in different cities of Pakistan

A line graph of grade point average



Scatter plot of GPA and the Starting Salary