Objectives

- Create the basic types of pie charts and bar graphs
- Interpret data given in pie charts and bar graphs
- Understand the purposes and limitations of graphs
- Understand the various elements of graphs

Sep 11-3:24 PM

Graphical Displays of Data: Pie Charts and Bar Graphs

The Value of Graphs

Trade-off:

- lose sight of the individual observations (the raw data)
- able to see a representation of the totality of observations

A well-designed graph gives our visual processing system the kind of image it processes best, a picture.

An examination of publications such asTime, USA Today, The Wall Street Journal, Scientific American, or Forbes provides convincing evidence of the frequent and beneficial usage of these graphical display techniques.

Sep 11-3:32 PM

Selecting and creating graphical displays requires a certain amount of artistic judgment.

Fortunately, the development of graphics software has made the creation of sophisticated graphs quite easy.

Several types of graphs and tabular displays are discussed in this chapter:

- bar charts (2-D and 3-D)
- pie charts
- line charts
- stem-and-leaf diagrams
- several types of frequency distributions
- histograms

Sep 11-6:40 PM

Bar Charts

Bar charts are often used to illustrate a frequency distribution for qualitative data.

Definition

Definition

The **bar chart** is a simple graphical display in which the length of each bar corresponds to the number of observations in a category.

Bar charts are valuable as presentation tools and are especially effective at **reinforcing differences in magnitudes**, since they permit the visual comparison of data by displaying the magnitude of each category by a vertical or horizontal bar. Figure 3.1 is a bar chart constructed from a housing type distribution for students in a statistics class.

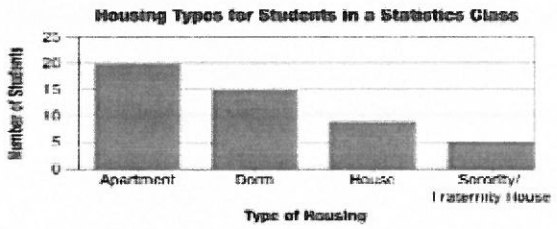


Figure 3.1

The Aesthetics of Bar Chart Construction

Certain conventions improve the quality and effectiveness of charts:

- Bar charts can be constructed **horizontally or vertically**. Customarily, horizontal orientation is used for categories that are descriptively labeled, and vertical (or columnar) orientation is used for categories that are numerical.
- If the categories have **some associated order,** they maintain that order in the bar chart. Otherwise, the categories may be listed alphabetically, in either ascending or descending order, or in some other pattern related to the nature of the data.
- **Miscellaneous or "other" categories should be listed at the bottom** of the chart (if oriented horizontally) or at the far right (if oriented vertically).

Sep 11-6:50 PM

> The difference in bar length is the principal visual feature in comparing differences in category amounts. Consequently, **scales for the axes should be chosen that will most effectively allow for the desired comparisons**. Unless there is a good reason, the axis used to measure the bars should *start at zero.* Otherwise the axis can be stretched to exaggerate differences in the bar lengths. For example, suppose the data in the following table were plotted.

Table 3.19 – Sales Performance	
Salesperson	Total Sales (Thousands of Dollars)
Susan	187
William	201
Beth	207
Rob	193

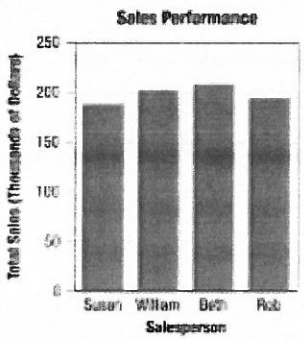


Figure 3.2

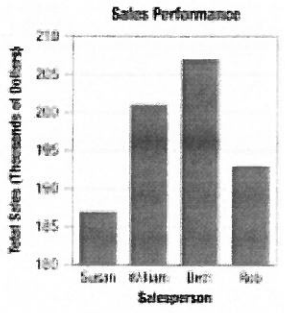


Figure 3.3

When you see an axis that does not start at zero, you should be a bit skeptical as to the conclusions the author intends for you to make.

Sep 11-6:50 PM

- Bar widths should be chosen that are visually pleasing and should not be allowed to vary within a particular chart.
- The spacing between bars can dramatically affect the perception of the graph. Spacing should be set at approximately one-half the width of a bar.
- Gridlines extended into the body of the chart are often useful and may be included if deemed helpful. Study Figure 3.4 and Figure 3.5.

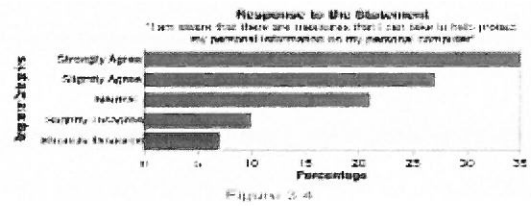


Figure 3.4

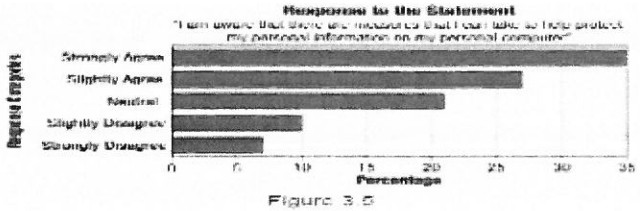


Figure 3.5

- Labels should be provided for each bar (category) and for each axis.
- Notes on sources of data or other footnotes should be given below the chart.
- A title for the chart that describes the data being presented should be added above the chart.

Stacked Bar Charts

Stacked bar charts are an interesting variation on the standard bar chart.

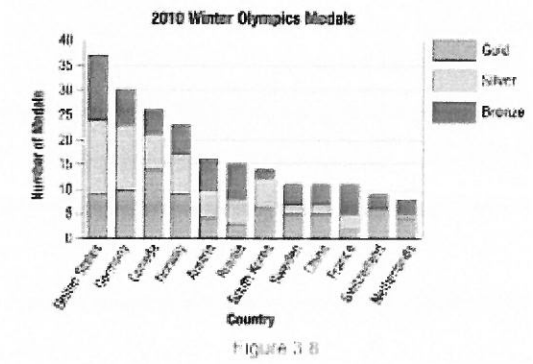


Figure 3.8

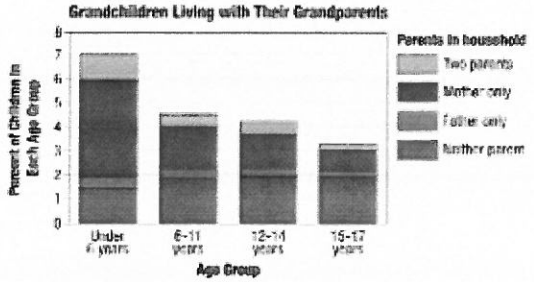


Figure 3.9

Stacked bar charts are useful when there are three components to the data (in this case the percentage of children, the age group, and the parents in the household).

3-D Bar Charts

Another interesting way of looking at the Olympic medal data is to plot a three-dimensional bar chart (see Figure 3.10). Using this graph, the totals for each type of medal are graphed for each country.

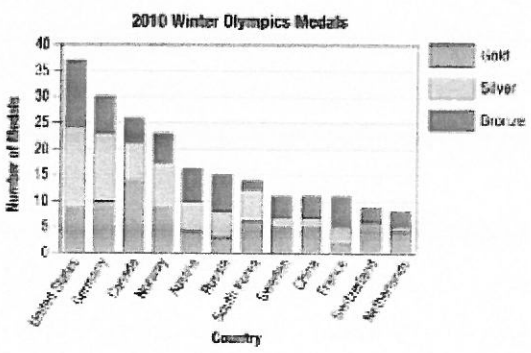


Figure 3.8

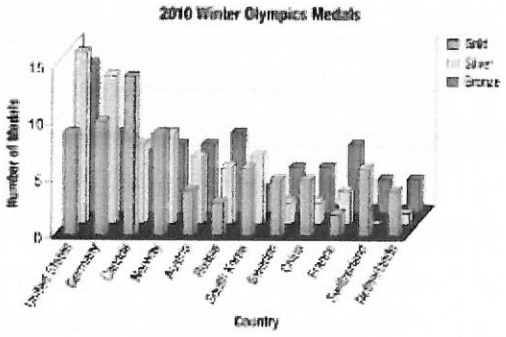
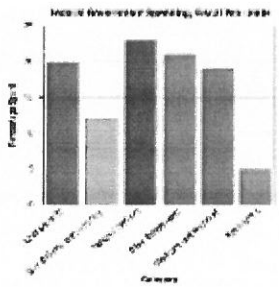


Figure 3.10

Pie Charts

We have just seen how bar charts are used as a means of expressing frequency distributions. Pie charts can perform the same function.

Table 3.20 – Percentage Spent by the Federal Government in 2009	
Category	Percentage Spent
Social Security	20%
Non-Defense Discretionary	12%
National Defense	20%
Other Entitlements	21%
Medicare and Medicaid	19%
Net Interest	5%



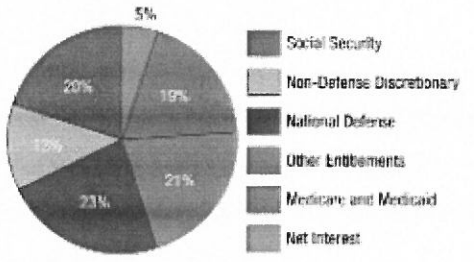
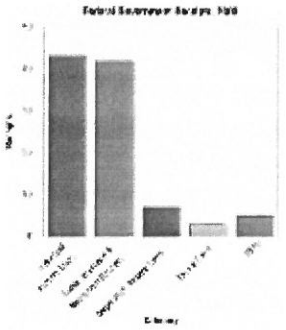
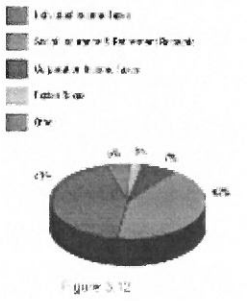


Figure 3.11





Objectives

- Construct dot plots from the data given
- Construct frequency polygons from the data given
- Construct histograms from the data given
- Construct ogives from the data given
- Construct stem and leaf plots from the data given
- Interpret information shown in line graphs, histograms, frequency polygons, ogives, and stem and leaf plots
- Read and interpret the information shown in dot plots

Sep 11-8:23 PM

Definition

A **histogram** is a bar graph of a frequency or relative frequency distribution in which the height of each bar corresponds to the frequency or relative frequency of each class.

Once the frequency distribution has been calculated, all the information necessary for plotting a histogram is available. In Figure 3.13, the histogram is created from the frequency distribution of the revenue data from the top 100 companies in the Fortune 500 in the table below.

Sep 11-8:23 PM

Frequency Distribution	
Revenue (Millions of Dollars)	Number of Companies
0 to 40	50
41 to 81	30
82 to 122	14
123 to 163	3
164 to 204	1
205 to 245	0
246 to 286	1
287 to 327	0
328 to 368	0
369 to 409	1

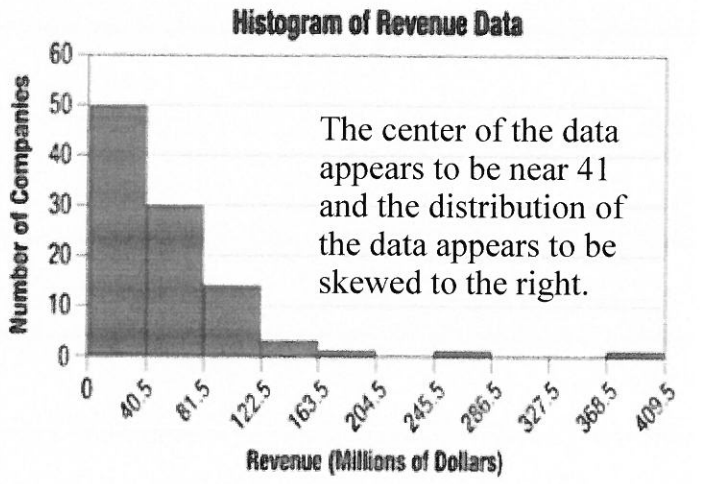


Figure 3.13

A symmetric distribution is one in which if a line were drawn down the middle of the distribution, the two sides would mirror each other. A skewed distribution is represented by a group of observations that is not equal on both sides (sometimes called asymmetric).

Sep 11-8:23 PM

You will see many histograms throughout this course. When we look at a histogram, what features are important?

1. Is the distribution symmetric or skewed (one tail is longer than the other)?

2. Is it bell-shaped?

3. Does the distribution have several peaks or modes?

4. Where is the center of the distribution?

5. Are there outliers (data values that are very different from the others)?

Sep 11-8:23 PM

Procedure: Constructing a Histogram

1. Determine the number of classes (or intervals) for the histogram. There isn't a rule of thumb for the number of classes to use, and most of the time the number of classes is determined by trial and error. However, a good starting point could be to use $\sqrt{n}$ or $\sqrt[3]{2n}$.

2. Determine the largest and the smallest observations. Depending on the size of the data set, it may be a good idea to sort the observations from smallest to largest. This will also make it easier later when trying to determine the interval in which an observation falls.

3. Calculate the difference between the smallest and largest observations by subtracting the smallest observation from the largest observation. (Note that this measure is called the range, which will be discussed further in a later chapter.)

4. Calculate the class width, using the following formula.

$$\text{Class Width} = \frac{\text{Largest Value} - \text{Smallest Value}}{\text{Number of Classes}}$$

Sep 11-8:23 PM

5. Calculate the limits of the class intervals. The rule of thumb for determining the class intervals is to choose the lower class limit of the first class so that it is either the minimum data value or a smaller number, so that the smallest observation falls in the first class. The first lower limit should have the same number of decimal places as the largest number of decimal places in the data. After choosing the lower limit of the first class, add the class width to it to find the lower limit of the second class. Continue this pattern until you have the desired number of lower class limits. The upper limit of each class is determined such that the classes do not overlap. If, after creating your classes, there are any data values that fall outside the class limits, you must adjust either the class width or the choice for the first lower class limit.

6. Calculate the class boundaries. A class boundary is the value that lies halfway between the upper limit of one class and the lower limit of the next class. To find a class boundary, add the upper limit of one class to the lower limit of the next class and divide by 2. After finding one class boundary, add (or subtract) the class width to find the next class boundary.

7. Tally each observation to determine in which interval it should fall. This will give you the class frequency (denoted by f_i), which is the number of observations in each interval.

8. All of the information for a histogram can be displayed in a frequency distribution table or a relative frequency distribution table, as can be seen in the following example.

9. Construct the histogram with the class boundaries on the horizontal axis and the frequency (or relative frequency) on the vertical axis. The heights of the bars will represent the class frequencies (or relative frequencies). Label the horizontal axis using either the class boundaries or class midpoints.

Example 3.8

Suppose you are given a random sample of 28 applicants who took an examination designed to measure their aptitude for a job in sales. The following scores (measured in percentages) were obtained. Note that the observations are already sorted from smallest to largest.

Table 3.22 – Aptitude Scores (%)

38	49	53	56	58	58	60
62	66	67	69	69	71	74
75	76	77	77	77	78	78
81	82	83	84	87	88	88

Construct a relative frequency distribution table for the test scores using 6 class intervals, and construct a histogram.

Table 3.22 – Aptitude Scores (%)						
38	49	53	56	58	58	60
62	66	67	69	69	71	74
75	76	77	77	77	78	78
81	82	83	84	87	88	88

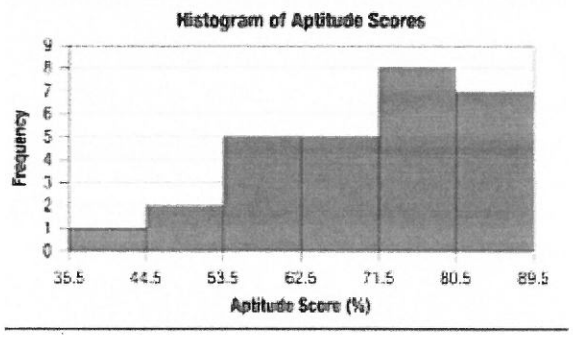
Solution

To solve this problem, we will work through the 9-step procedure for constructing a histogram.

1. First, we need to determine the number of classes for the histogram. We are given that we want to construct a histogram using 5 intervals, so there will be 5 classes.
2. The second step is to identify the largest and the smallest observations in the data set.
3. Next, we need to calculate the difference between the smallest and the largest observations.
4. Now we need to calculate the class width.
5. The next step is to construct the class intervals.
6. Once we have calculated the class limits, we calculate the class boundaries.
7. Tally each observation to determine the interval in which each observation falls.
8. The relative frequency distribution table provides a summary of the data and allows the histogram to be constructed.

Table 3.26 – Relative Frequency Distribution

Class Limits	Class Boundaries	Frequency (f)	Relative Frequency
38–44	35.5–44.5	1	0.0357
45–53	44.5–53.5	2	0.0714
54–62	53.5–62.5	5	0.1786
63–71	62.5–71.5	5	0.1786
72–80	71.5–80.5	8	0.2857
81–89	80.5–89.5	7	0.2500
	Total	28	1.0000



The Stem-and-Leaf Display

The stem-and-leaf display is a hybrid graphical method. The display is similar to a histogram, but the data remain visible to the user. It is one of the few graphical methods in which the raw data are not lost in the construction of the graph. As the name implies, there is a "stem" to which "leaves" will be attached in some pattern.

Consider the following data: 97, 99, 108, 110, and 111. If we are interested in the variation of the last digit, the stems and leaves are as shown in Table 3.27 and displayed in Figure 3.14. The leaves in this case are the ones digits and the stems are the tens digits. All of the data values that have common stems are grouped together, and their leaves branch out from the common stem.

Table 3.27 – Data, Stems, and Leaves

Data	Stem	Leaf
97	09	7
99	09	9
108	10	8
110	11	0
111	11	1

Stem-and-Leaf Display

Stem	Leaf
09	7 9
10	8
11	0 1

Figure 3.14

Stem-and-Leaf Display

Stem	Leaf
0	97 99
1	08 10 11

Figure 3.15

Sep 11-8:36 PM

The Ordered Array

An ordered array is a listing of all the data in either increasing or decreasing magnitude. Data listed in increasing order are said to be listed in rank order. If listed in decreasing order, they are listed in reverse rank order. Listing the data in an ordered way can be very helpful. It allows you to scan the data quickly for the largest and smallest values, for large gaps in the data, and for concentrations or clusters of values

Table 3.30 – Ages (Raw)

32	21	24	19	61	18	18	16	16	35	39	17	22
21	60	18	53	18	57	63	28	20	29	35	45	

Table 3.31 – Ages (Ordered)

16	16	17	18	18	18	18	19	20	21	21	22	24
28	29	32	35	35	39	45	53	57	60	61	63	

Sep 11-8:36 PM

Dot Plots

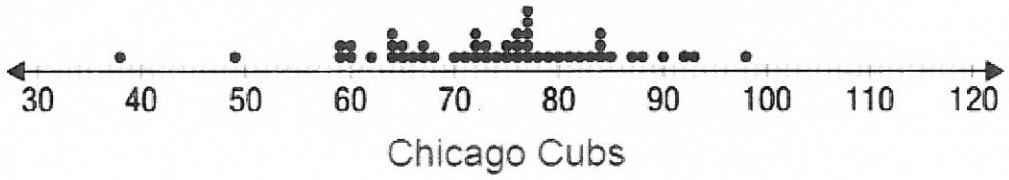
A dot plot is a graph where each data value is plotted as a point (or a dot) above a horizontal axis. If there are multiple entries of the same data value, they are plotted one above another.

Example 3.11

The following table contains the number of wins by baseball's Chicago Cubs for a 50-year period. Use this data to construct a dot plot.

Table 3.32 – Wins by the Chicago Cubs (1953–2002)

67	68	78	76	73	81	85	87	59	62
88	70	77	70	38	75	83	59	64	60
65	73	77	77	64	75	84	72	68	79
67	49	93	90	80	66	92	76	74	64
90	84	77	71	78	77	84	88	72	66



Plotting Time Series Data

Recall that the science of statistics is divided into two categories: descriptive statistics and inferential statistics. Fundamental to the concept of statistical inference is the notion of population–the total collection of measurements. Time series data originate as measurements usually taken from some process over equally spaced intervals of time.

Table 3.34 – U.S. Census Data

Year	Population (Millions)	Year	Population (Millions)
1790	3.9	1900	76.2
1800	5.3	1910	92.2
1810	7.2	1920	106.0
1820	9.6	1930	123.2
1830	12.9	1940	132.2
1840	17.1	1950	151.3
1850	23.2	1960	179.3
1860	31.4	1970	203.3
1870	38.6	1980	226.5
1880	50.2	1990	248.7
1890	63.0	2000	281.4
		2010	308.7

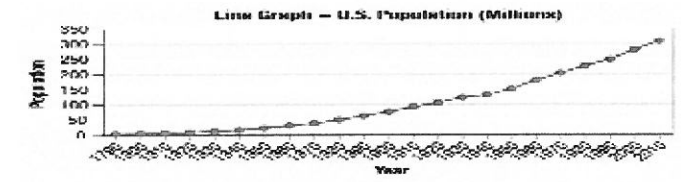


Figure 3.17

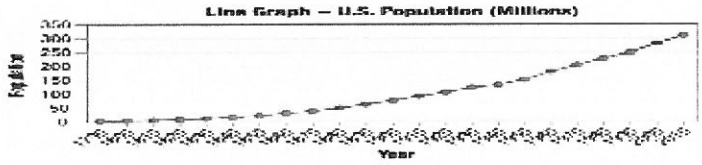
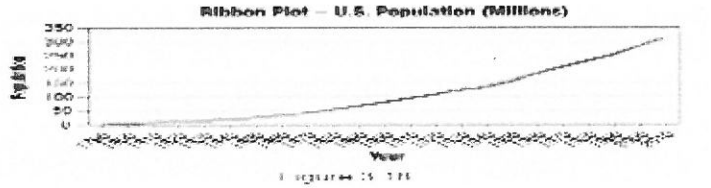


Figure 3.17



Sep 11-8:36 PM

CAREERS

Health informatics is one of the fastest growing career fields. By combining their skills in public health and information, health informatics professionals can undertake many unique and important roles.

The Master of Science in Information (MSI)

The MSI is a professional degree which prepares students for emerging careers that meet the rapidly growing information-management needs of an increasingly interconnected world. As businesses and society grapple with the challenges and opportunities of the digital age, information professionals play a crucial role in analyzing, systematizing, and evaluating the massive resources generated by the digital revolution. At Michigan, we train students to be leaders and agents of change in a field that is evolving at unprecedented speed.

10. Data Modeler

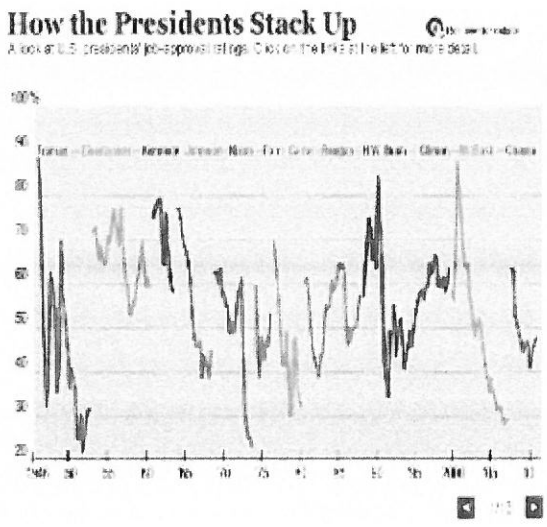
Another position that translates poorly without jargon, these IT professionals create data designs and define relationships between data fields, according to TBS. Since any company's data is vital, it's modeling needs to work perfectly – a more complex task as reliance on computers grows.

Education: Bachelor's degree in computer science, mathematics or IT – plus on-the-job experience, says TBS.

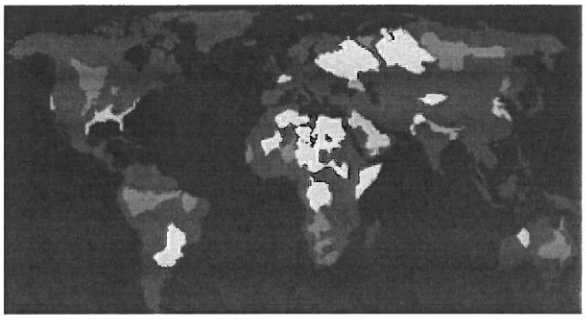
Salary: A hefty $103,000, according to TBS.

Sep 11-6:58 PM

An examination of publications such asTime, USA Today, The Wall Street Journal, Scientific American, or Forbes provides convincing evidence of the frequent and beneficial usage of these graphical display techniques.



http://online.wsj.com/public/resources/documents/info-presapp0605-31.html



http://www.usatoday.com/pages/interactives/groundwater/

Looking Ahead

Data analytics specialists are in high demand, as cited in the Globe and Mail article "*The fastest growing job market you never heard of*".

User experience designer, information architect, data security analyst are not only growing career fields, but among the higher paying ones too, according to a study done by the staffing firms Robert Half and The Creative Group. And this is a worldwide trend too.

According to a recent SAS UK research, "*demand for big data specialist will grow by 243% in the next 5 years*". Similarly, in the U.S., by 2018, there will be a shortage of "*1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions*" according to McKinsey Global Institute.

The large baby boomer generation in traditional libraries will be retiring in the next 5 to 10 years, which will create the need for new librarians, and library managers, in all types of libraries and roles. Employment in special, corporate libraries, in settings such as financial institutions, consulting firms, government, law firms, and technology companies, is expected to grow at a faster than average pace. This rapidly advancing field is anticipated to offer many employment opportunities for information professionals.

https://www.ted.com/talks/
hans_rosling_shows_the_best_stats_you_ve_ever_seen

Sep 11-6:47 PM