ST101 Unit 1: Visualizing Relationships in Data

Contents

Welcome

Welcome to Statistics 101. Let’s start with a teaser. This is a challenging teaser, and it may provoke you.

I believe you should be unhappy. Not because our class is bad, because I will prove in a moment that you are unpopular. The reason we’re doing this is to show how deep statistics is, and how we can easily fool ourselves.

For simplicity, let's say that there are two types of people, type A and type B. Type A are popular. They have 80 friends. Type B are less popular. They only have 20 friends.

Now, you may now say that I don't know which type you are. We will calculate the expected, or average number of friends. To do this, we assume that half of the people are of type A and the other half are of type B.

Average Friends Quiz

What is the average number of friends you have, if you have a 50% chance of being a type A and a 50% chance of being a type B?

Expected Friend Type Quiz

Now, each of your friends on Facebook or Google+ will have either 80 friends or 20 friends. If you pick one friend at random, what is the chance that you have picked a Type A friend? What is the chance that they are type B? (The two numbers should sum to 1).

Unpopular Quiz

Let’s get back to the real question. How many friends should you expect the friend that you picked to have?

Course Overview

Most of the material we will cover in this class is very basic. It is the first class you would have in college if you're not a statistics major. We’ll teach you how to visualize data, how to summarize it, how to run tests, and even how to find trends. But there are also a few challenging nuggets in there.

The challenges are optional, and they’re clearly marked as optional. You will prove some theorems along the way, and – most importantly – you’ll get the chance to program the things you've learned (using the Python programming language).

Again, programming is optional. You don’t need to have a programming background to do this course. But we would suggest you give it a try. Many people learn the material much better by programming than any other way.

Looking at Data

The basis of statistics is that the world is full of data, and we have to make decisions. Statistics comes to our rescue. It takes data and turns it into information that we can use to make decisions. Whatever field you are in, the chances are that it is driven by data. Statistics is important to know and to understand. It's universal, useful, and Sebastian promises it is fun too!

Valuing Houses

One of the standard problems that people study in statistics has to do with purchasing decisions. Suppose you want to buy a house. There are houses of various sizes, but you really like one particular house. This house has a specific asking price, let's say $92,000.00. You want to know whether this is too much, or perhaps too little?

In statistics, the way to find out is by looking at data. Let's assume there's a database of previous house sales in the same neighbourhood. For simplicity, we'll assume we know two things, the size of the home and the sale price.

Size (ft^2)	Cost ($)
1400	112,000
2400	192,000
1800	144,000
1900	152,000
1300	104,000
1100	88,000

Valuing Houses Quiz

The house you wish to purchase is 1300 square feet in size. How much should you expect to pay?

Valuing Houses Quiz 2

How much should you expect to pay for a house with 1800 square feet?

Valuing Houses Quiz 3

What if the house you want to buy is 2100 square feet?

Valuing Houses Quiz 4

What about 1500 square feet?

Valuing Houses Quiz 5

What is the cost of a home per square foot?

Scatter-Plots

Most Important Part Quiz

What do think is the most important thing that a statistics person does?

- Look at data
- Program computers
- Run statistics
- Eat pizza

Linear Relationship 1 Quiz

Here is another data set:

Size (ft^2)	Cost ($)
1400	98,000
2400	168,000
1800	126,000
1900	133,000
1400	91,000
1100	77,000

Is there a fixed cost per square foot for this data set?

Linear Relationship 2 Quiz

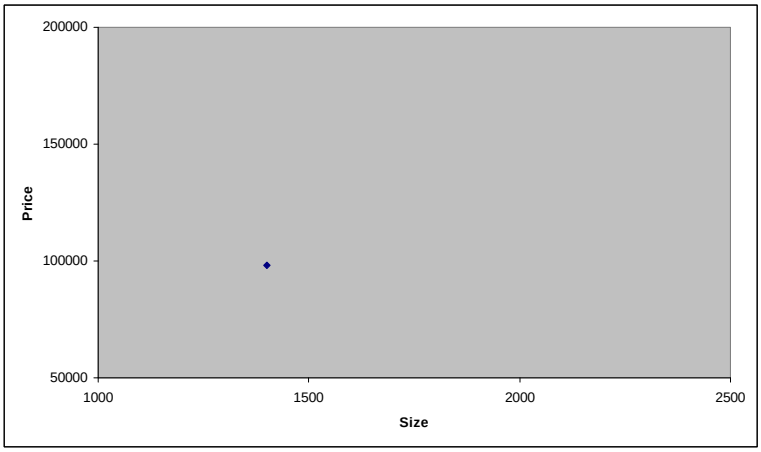
What if I change the second 1400 square foot house to 1300 square feet:

Size (ft^2)	Cost ($)
1400	98,000
2400	168,000
1800	126,000
1900	133,000
1400	91,000
1100	77,000

Is there a fixed cost per square foot for this data set now?

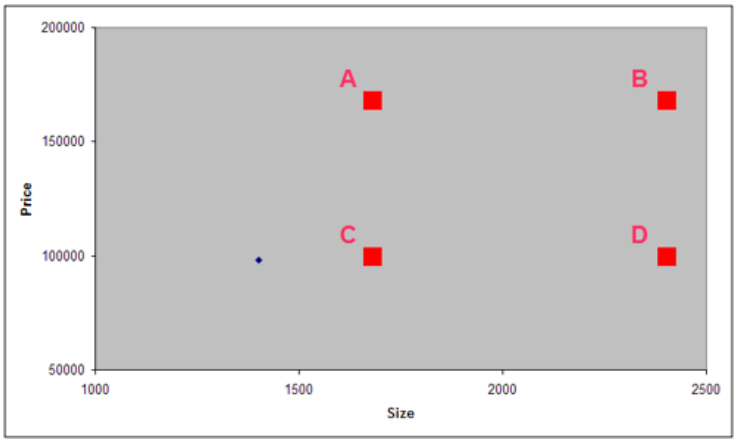
As we have seen, data can carry a lot of information. There is a trick called a scatter-plot that you can use to visualise the data. Take a pencil and a piece of paper and arrange the data in a graph where the x axis is house size, and the y axis is the price.

In a scatter plot, each data item becomes a dot on the graph. The first house would appear on the graph as follows:



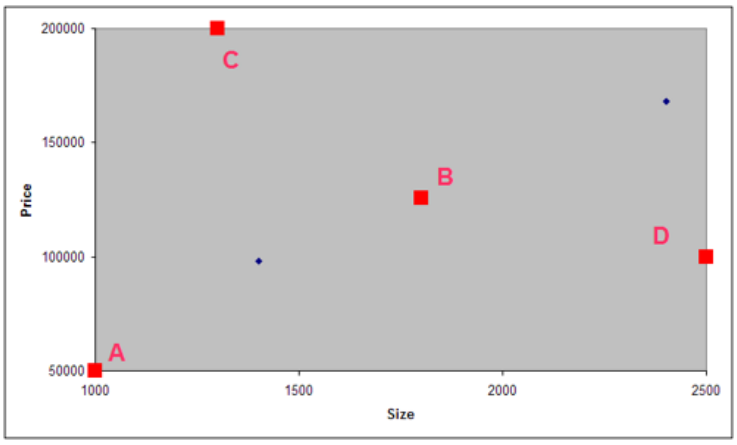
Scatter Plot Quiz

Where would the second house be plotted?

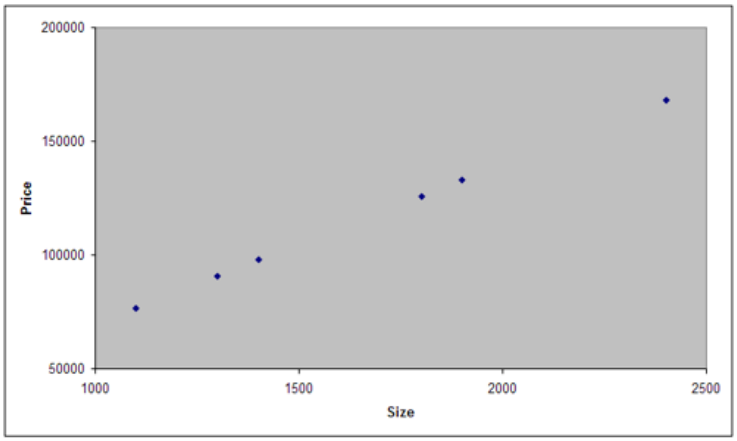


Picking Points Quiz

What about the third house?



Now, we chose a 2-dimensional list to make for a convenient 2-dimensional scatter-plot. These are the most popular scatter plots because surfaces like paper are 2-D. When we add the remaining data points, we get:



This is a nice scatter-plot that allows us to draw a straight line through all the points. When this happens, and there's a relationship between the data that is governed by a straight line we call the data **linear**. Linearity is fairly rare in statistics. More often you will find deviations, because the size of a house is not the only factor that determines its cost (or perhaps also because most of us are bad negotiators!). When a data set is linear, it is really easy to predict the prices of houses in between, just be looking at the data. We're doing what a statistician ought to do.

Make Your Own Quiz

Plot the following data as a scatterplot. Is the relationship between price and size linear?

Size (ft^2)	Cost ($)
1700	51,000
2100	63,000
1900	57,000
1300	39,000
1600	48,000
2200	66,000

Fixed Price Quiz

Do you think there's a fixed price per square foot for this data?

Price Per Square Foot Quiz

Now do you think there's a fixed price per square foot for this data?

Size (ft^2)	Cost ($)
1700	53,000
2100	65,000
1900	59,000
1300	41,000
1600	50,000
2200	68,000

Make Your Own Quiz 2

Plot the following data. Is the data linear? Can you fit a line through the scatter plot data?

Size (ft^2)	Cost ($)
1700	53,000
2100	65,000
1900	59,000
1300	41,000
1600	50,000
2200	68,000

Find The Constant Quiz

That is a surprising result. Even though there's no fixed cost per square foot, the relationship between the data is linear. Actually, in this case the price per square foot is linear, plus or minus a constant dollar amount. Can you work out the amounts?

Price = $_____ per square foot x size + $_____

Is it Linear? Quiz

Plot the modified data below. Is the relationship between the data linear?

Size (ft^2)	Cost ($)
1700	53,000
2100	44,000
1900	59,000
1300	82,000
1600	50,000
2200	68,000

Congratulations

So, now we know a lot about scatter plots. They tend to be 2-dimensional, and a simple eye-ball of the data can tell us a lot the relationship of one variable to another.

Scatter-plots aren't great when there is what is called "noise" in the data. This happens when the data deviates from expectation in some random, noisy way. Next, we'll look at another simple plotting technique called bar charts that address the issue of noisy data by grouping data points into a single cumulative bar.

Bar Charts

In this section we're going to look at bar charts. These are a common statistical data visualization tool.

Checking Linearity Quiz

Let's look at our housing data again. This time the data is ordered by increasing house size. Is this data linear?

Size (ft^2)	Cost ($)
1300	88,000
1400	72,000
1600	94,000
1900	86,000
2100	112,000
2300	98,000

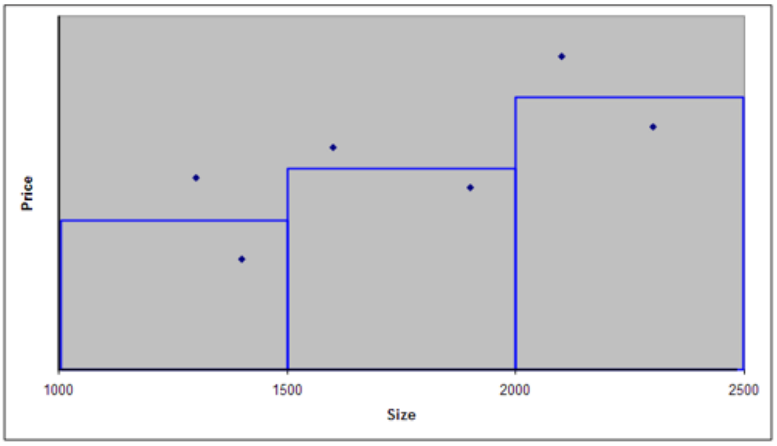
Interpolation Quiz

If we now ask how much you should pay for a 2200 square foot house, using the interpolation method we learned earlier, what figure would you get? Do you trust that number?

There is good reason not to trust this value. The cost of a 2300 square-foot house is less than the 2100 square-foot house. These deviations from the linear relationship are called **noise**. This is the term that statisticians use. Maybe one house has a great view, while another is an old house. Perhaps a third is on the coast, or maybe one needs a new kitchen. There are a whole range of possible factors that effect the cost over and above the size of the property.

If these factors aren't included in the data, a statistician will call it **random noise**. Bar charts are one way to alleviate the problem.

In a bar chart, we take the raw data and pool it together into ‘bands’. For example, in our house data, we may group all the data for house sizes between 1000 and 1500 square-feet into one bar. Then group the data for house sizes between 1500 and 2000 square-feet into another bar, and so on:



Grouping Data Quiz

What should the height of the first bar (from 1000 to 1500 square-feet) be in dollars?

Grouping Data Quiz 2

What about the dollar values of the heights of the second and third bars?

Bar Charts

When we look at the bar chart, we will see that it is a much finer representation of the data. Pooling multiple data points together to form a single bar, can give a much clearer picture of the dependence of cost on size. While the bar chart doesn’t show the linear relationship in the same way as the scatter-plot (actually, in this case the relationship is non-linear), it really gives a clear sense that, as house size increases, the cost increases. Something that may not have been obvious from just looking at the individual data points.

The bar chart lets us pool groups of data together into single bars and so understand global trends. Now, these global trends might not be that important if you only have six data points, but imagine that you have 60,000 data points. In this case, small variations in individual data points may not tell us much, but the bar chart can really help us to understand the data.

One of the jobs of the statistician is to use cumulative tools, such as bar graphs, to gain an understanding of the underlying data.

Histograms

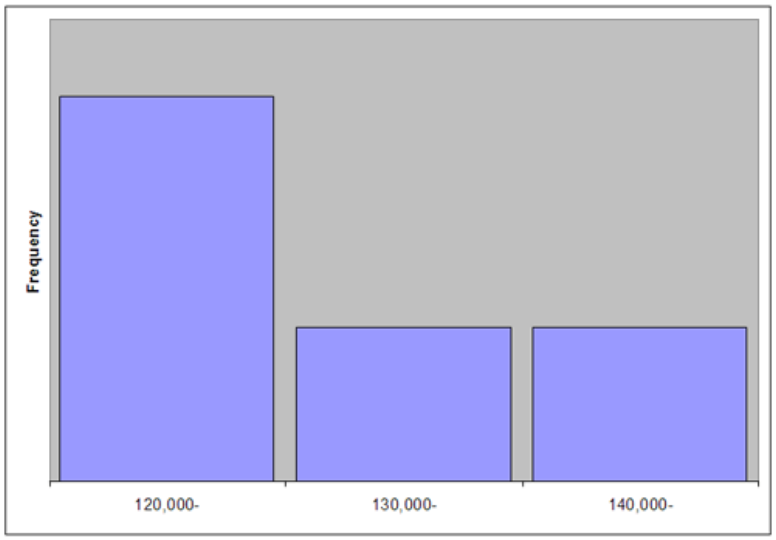
Now, we are going to introduce histograms as a special case of the bar chart.

The key difference is that the bar charts that we have discussed so far have dealt with 2-dimensional data. Histograms only consider 1-dimensional data. Let's consider an example.

Let's suppose that we asked a group of software engineers how much they earn, and got the following responses:

$132,754
$137,192
$122,177
$147,121
$143,000
$126,010
$129,200
$124,312
$128,132

For the histogram, we are going to create a bar chart that is only concerned with frequency. This is basically a count that groups the salaries into a series of buckets, say from $120,000 to $130,000, from $130,000 to $140,000, and over $140,000.

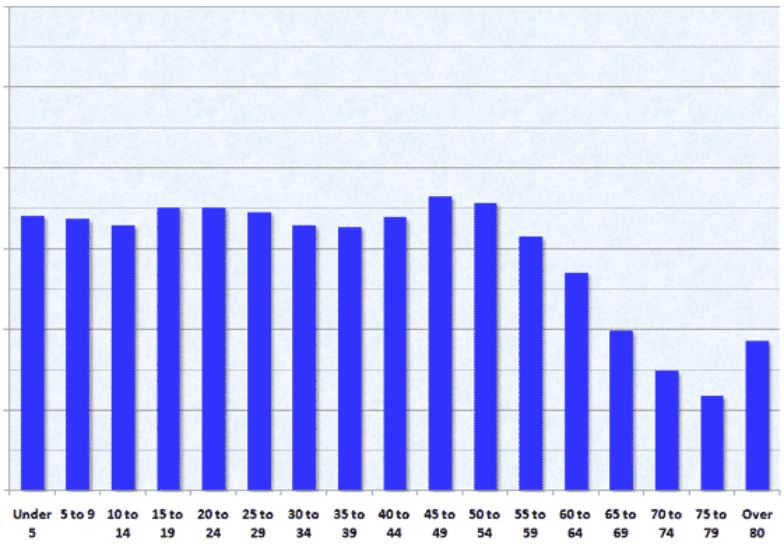


Histogram Quiz

What is the frequency count for the salaries that fall into the three brackets?

Age Distribution Quiz

A well-known histogram can be obtained by looking at age distributions. For the USA, the distribution looks something like this:



Let's create a rather simplified histogram, looking at people between 0 and 40 using the following data set:

21, 17, 9, 27, 35, 4, 12, 12, 32, 14, 38, 9, 19, 22, 21, 14, 3, 8, 31, 15, 33, 29

Group the data into the ranges:

0 – 10, 11 – 20, 21 – 30, and 31 – 40.

What are the heights of the bars for the four ranges?

Summary

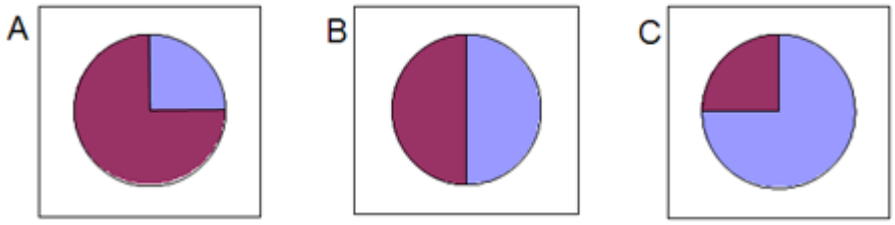
In this unit we learned about bar charts and histograms. Both use vertical bars, and both aggregate data. The big difference is that bar charts are defined over 2-D data, one dimension applied to the x axis and the other to the y axis. Histograms only apply to 1-D data, and the y axis becomes the count of that data.

Pie Charts

Most of us have seen pie charts before. In statistics, we use pie charts to visualise data. Specifically, relative data, and we will see what that means in a moment.

Voting Quiz 1

Let's say that there is an election and there are just two parties. Both parties are getting the same number of votes, i.e. 50%. Which of these pie charts reflect the outcome of the election?



Voting Quiz 2

Now, we said that pie charts are good for relative data. Suppose Party A got 724,000 votes and Party B got 181,000 votes. What percentage of the vote did Part A get?

Voting Quiz 3

Now, given this, which of these charts most closely resembles the election result?

Relative Data Quiz

Now, given that we know that the distribution of the votes was 80% to 20%, if we know that 23,000 people voted for party B, how many people voted for party A?

So, a remarkable property of pie charts is that they are invariant to the actual numbers of votes. What it actually depicts is the ***relative*** numbers of votes. In this case, it show

that Party A got many more votes than Party B. It shows this graphically, so you can see this without having to study the actual number of votes cast.

Pick the Breaks Quiz

You are studying a Udacity course. You find the following age distribution among the students on the class:

Age	# Students
13 – 19	12,000
20 – 32	96,000
33 - 999	36,000

Construct the pie chart for this data.

Build a chart Quiz

Let's build another pie chart. Let's assume this time that our election ran with four parties. This was the result:

Party A	175,000
Party B	50,000
Party C	25,000
Party D	50,000

Construct the pie chart for this data.

Inferring Counts Quiz

In another election, the distribution of votes between the four parties remained unchanged, but the total number of votes was 240,000. How many votes were cast for each of the parties?

Party A	140,000
Party B	40,000
Party C	20,000
Party D	40,000

Now, the chart tells us nothing about the absolute number of votes cast, but it does tell us a lot about the distribution of the votes. We can easily see that A is the dominant party with more than 50% of the votes cast.

Summary

So we just learned about pie charts. We learned that they are great for relative data, and they're wonderful for comparing which slice of the pie is biggest. We will look at relative data again later with a case study about gender discrimination in college admissions, using a study originally performed at UC Berkeley in California.

Programming Charts (Optional)

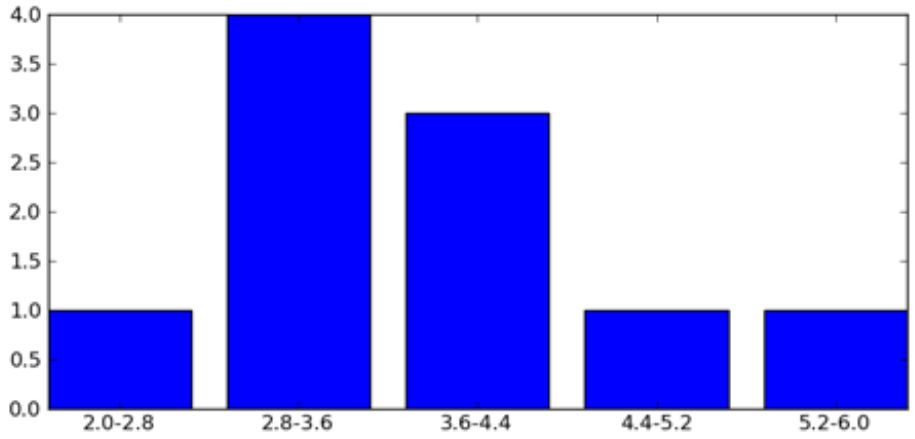
In this section, you get the chance to experience how to visualise data first hand. We will provide the data, and you will plot it and then answer some simple questions about that data.

Here are three lines of code, 3 instructions that the computer has carried out for me:

```
from plotting import *
data = [3, 4, 2, 4, 3, 5, 3, 6, 4, 3]
histplot (data)
```

If you are not a programmer, you can ignore the first instruction. It tells the computer that we want to plot things and it will always be there. The second and third lines are the important ones.

The second line defines a data set. It is a list of 10 elements. The third line tells the computer to make a histogram plot of my data. We then get the result:



As you can see, in this case the range 2.8 to 3.6 was most frequent.

So, there are three things we need to tell the computer. We tell it we want to plot things. We define the data, and give it a name. We tell it the type of plot we want (in this case a histogram).

Plot Height Quiz

Here is the data about the height of a group of people:

> Height=[65.78, 71.52, 69.4, 68.22, 67.79, 68.7, 69.8, 70.01, 67.9, 66.78,
> 66.49, 67.62, 68.3, 67.12, 68.28, 71.09, 66.46, 68.65, 71.23, 67.13, 67.83,
> 68.88, 63.48, 68.42, 67.63, 67.21, 70.84, 67.49, 66.53, 65.44, 69.52, 65.81,
> 67.82, 70.6, 71.8, 69.21, 66.8, 67.66, 67.81, 64.05, 68.57, 65.18, 69.66, 67.97,
> 65.98, 68.67, 66.88, 67.7, 69.82, 69.09]

Plot the data as a histogram. What is the most frequent height?

Plot Weight Quiz

Let's try that again with the weights of the same group of people.

> Weight=[112.99, 136.49, 153.03, 142.34, 144.3, 123.3, 141.49, 136.46,
> 112.37, 120.67, 127.45, 114.14, 125.61, 122.46, 116.09, 140.0, 129.5, 142.97,
> 137.9, 124.04, 141.28, 143.54, 97.9, 129.5, 141.85, 129.72, 142.42, 131.55,
> 108.33, 113.89, 103.3, 120.75, 125.79, 136.22, 140.1, 128.75, 141.8, 121.23,
> 131.35, 106.71, 124.36, 124.86, 139.67, 137.37, 106.45, 128.76, 145.68, 116.82,
> 143.62, 134.93]

What is the most frequent weight?

Scatter Plot Quiz

Let's look at the combined data for height and weight. Create a scatter plot of the height of the individuals against their weight. The command to create a scatter plot is:

> scatterplot(Height, Weight)

Is the data:

- exactly linear
- approximately linear
- height and weight are completely unrelated

Barchart Quiz

Replace the scatter-plot with a bar-chart using the same two arguments as before.

Now the chart show clearly that as the height increases, so does the weight, but the differences between the heights of the bars suggests that the relationship is not exactly linear. In reality, of course we know that the relationship between height and weight in a population isn't linear, but for the sake of this exercise, it is the best of the three options we provided for you.

Wages Quiz

Write a line of code to print a scatterplot of Age on the horizontal axis against Wage on the vertical axis. What is the youngest age at which a person earns $267,000?

Wage Bar-Chart Quiz

Make a bar chart using the same data. Is the relationship between age and wage:

- exactly linear
- approximately linear
- there is no relationship between age and wage.

Most Common Age Quiz

Create a graph to answer the question, what is the most common age?

Conclusion

In this section we created the Python code to generate our own bar charts, histograms, and scatter plots. We can use this to study the data, and perhaps learn something about it.

Admissions Case Study

Statistics is not just a superficial field. It can be really deep. The problem that we will examine here derives from an actual study by the University of California Berkeley. They wanted to know whether their admissions procedure was gender biased. Sebastian looked at various admission statistics to understand whether their admission policies had a preference for a certain gender.

The numbers that we will be using are not the same as those from UC Berkeley. This is a simplified version of that problem, but the paradox that it illustrates is the same. It is called "Simpson's Paradox".

Admissions Quiz 1

Among male students, 900 applied for Major A and 450 were admitted. What is the acceptance rate as a percentage?

Admissions Quiz 2

In a second major, Major B, 100 male students applied and 10 were accepted. What is the acceptance rate as a percentage?

Admissions Quiz 3

The same statistic was run for female students. Females applied predominantly for Major B. There were 900 applications for major B, of whom 180 were admitted. Just 100 female students applied for Major A, of whom 80 were admitted.

What is the acceptance rate for Major A as a percentage for the female student population?

Admissions Quiz 4

What is the acceptance rate for Major B as a percentage for the female student population?

Gender bias quiz

So, just looking at these numbers for the two different majors, do we believe – in terms of the acceptance rate – that there is a gender bias? Is it in favour of male or female students?

Superficially, it appears that female students are favoured because for both majors, they have a better admission rate than the corresponding rate for male students. But what happens if we look at the admission statistics independently of the major?

Aggregation Quiz 1

A total of 1000 male students applied and 460 were admitted. What is the acceptance rate for male students across both majors?

Aggregation Quiz 2

Now do the same for female students. A total of 1000 students applied and 260 were admitted.

Gender Bias Revisited Quiz

So, across both majors, do we believe – in terms of the acceptance rate – that there is a gender bias? Is it in favour of male or female students?

Perhaps surprisingly, given our earlier findings, when we look at both majors together, we find that males have a much higher admissions rate than females. This is not made up. The actual numbers we are using may be made up, but this effect was actually observed University of California at Berkley many years ago.

Looking at majors individually, we find that in each major individually the acceptance rate for females trumps that of males, and yet when we look at the overall statistics we find the opposite. We haven't added anything. We just regrouped the data.

This example shows just how ambiguous statistics can be. In choosing how to graph your data, you can have a major impact on what people believe. A famous saying states “I never believe statistics that I didn't doctor myself”.

The key lesson here is that statistics can be deep and are often manipulated. You should always be sceptical of statistics, whether they are your own results or other peoples, and you really need to understand how raw data is turned into decisions or conclusions.

Answers

Average Friends Quiz

Now, I don't know which type you are, but there's a 50% chance that you're Type A, in which case you'll have 80 friends, and a 50% chance that you're Type B, and you'll have 20 friends. I can calculate your expected number of friends as:

(80 x 0.5) + (20 x 0.5) = 40 + 10 = 50 friends.

Expected Friend Type Quiz

Because Type A people are so much more popular, your chances of linking to a type A is 0.8. That chance that you picked a Type B friend is therefore 0.2.

Unpopular Quiz

Let's get back to the real question. How many friends should you expect the friend that you picked to have?

There is an 80% chance that you picked a Type A friend, who will have 80 friends. Similarly, there's a 20% chance that you picked a Type B friend who has 20 friends. This gives an expected number of friends:

(80 x 0.8) + (20 x 0.2) = 64 + 4 = 68 friends

You would only expect to have 50 friends, so this suggests that you are unpopular!

Valuing Houses Quiz

$104,000

Valuing Houses Quiz 2

$144,000

Valuing Houses Quiz 3

$168,000

Valuing Houses Quiz 4

$120,000

Valuing Houses Quiz 5

$80

Most Important Part Quiz

- **Look at data**
- Program computers
- Run statistics
- Eat pizza

Linear Relationship 1 Quiz

No. Two houses of the same size sold for different prices

Linear Relationship 2 Quiz

Yes. The cost per square feet is 70 dollars.

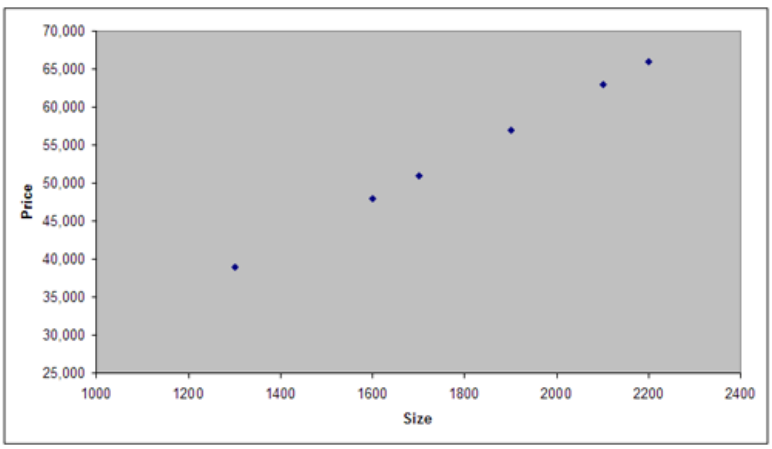
Scatter Plot Quiz

B

Picking Points Quiz

B

Make Your Own Quiz



Fixed Price Quiz

No

Price Per Square Foot Quiz

No. They're almost the same but not quite.

Make Your Own Quiz 2

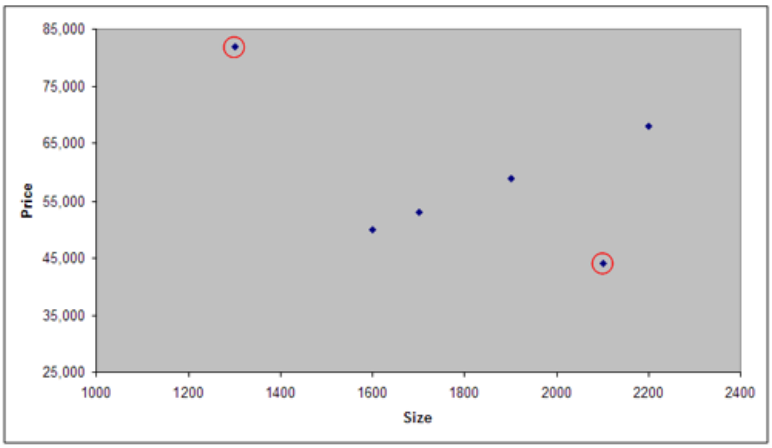
Yes. The data is linear.

Find The Constant Quiz

Price = $30 per square foot x size + $2000

Is it Linear? Quiz

No. The two highlighted values are '**outliers**'. We'll talk more about outliers later. There is no way to fit a linear function through all these data points.



Checking Linearity Quiz

No

Interpolation Quiz

$105,000

No, the value cannot be trusted.

Grouping Data Quiz

$80,000. The average (mean) of $88,000 and $72,000.

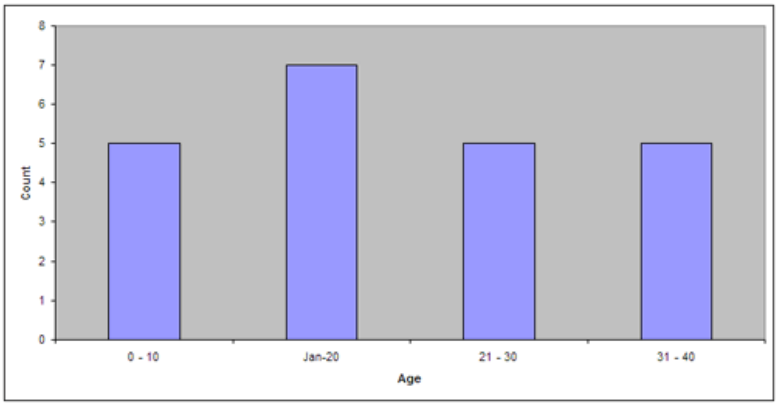
Grouping Data Quiz 2

$90,000 for the second bar and $105,000 for the third.

Histogram Quiz

$120,000 to $130,000	**5**
$130,000 to $140,000	**2**
$140,000 to $150,000	**2**

Age Distribution Quiz



Voting Quiz 1

B

Voting Quiz 2

80%

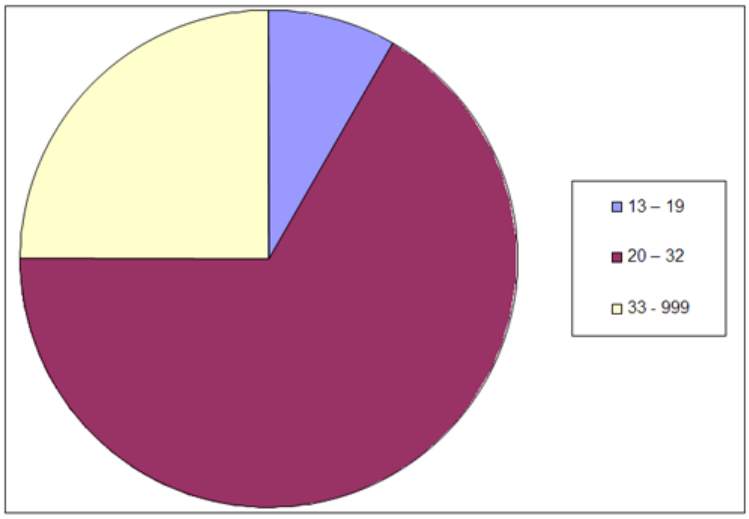
Voting Quiz 3

E

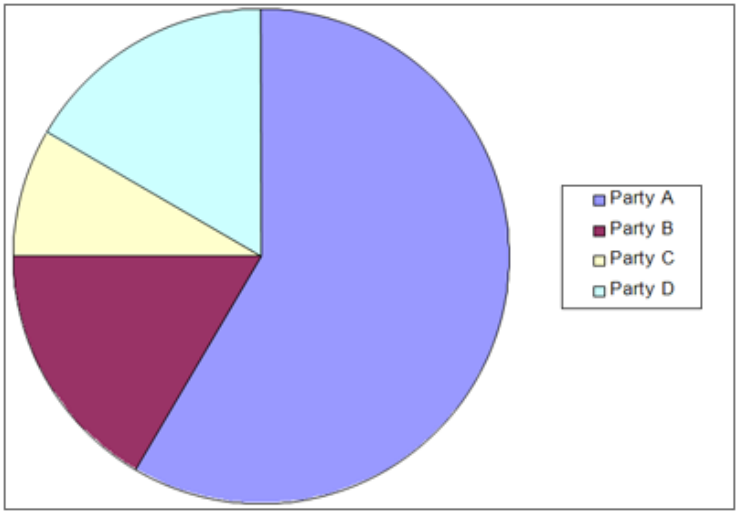
Relative Data Quiz

92000

Pick the Breaks Quiz



Build a chart Quiz



Inferring Counts Quiz

Party A	140,000
Party B	40,000
Party C	20,000
Party D	40,000

Plot Height Quiz

histplot (Height)

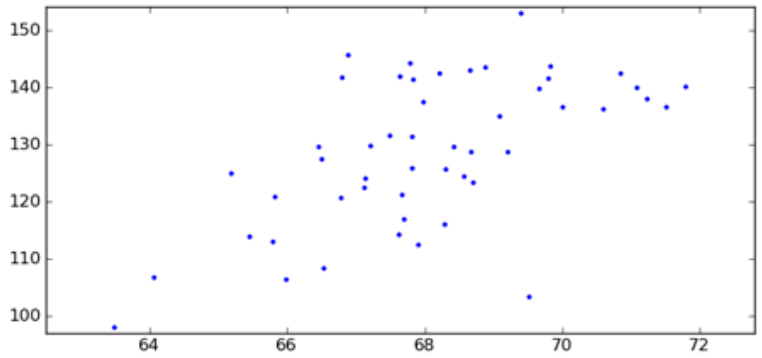
66 – 68 inches

Plot Weight Quiz

histplot (Weight)

119 – 130 lbs

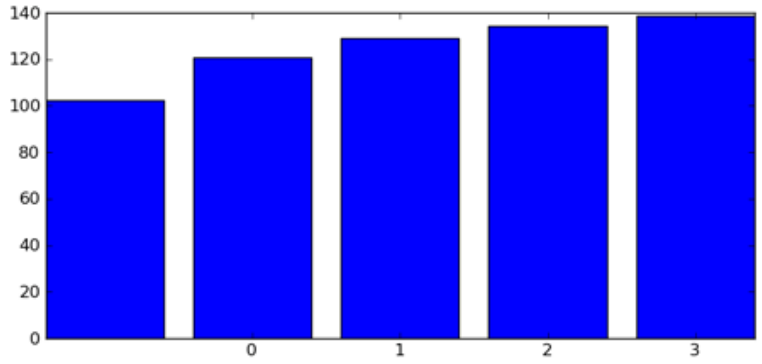
Scatter Plot Quiz



Approximately linear.

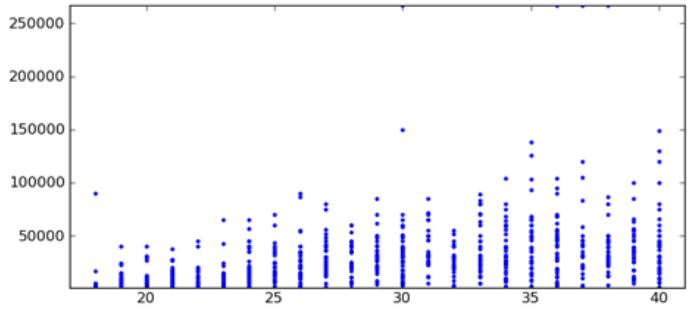
Barchart Quiz

barchart (Height, Weight)



Wages Quiz

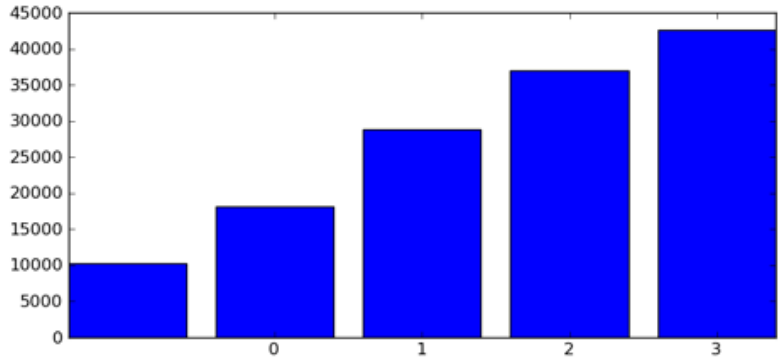
scatterplot (Age, Wage)



30.

Wage Bar-Chart Quiz

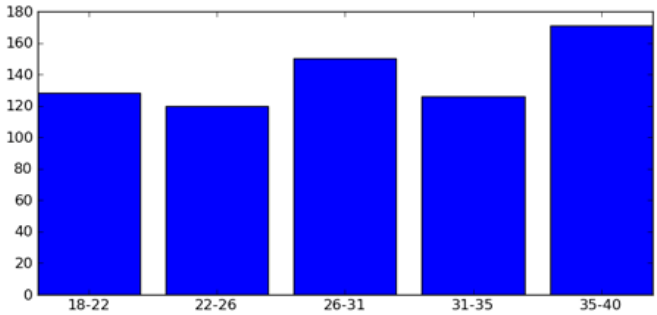
barchart (Age, Wage)



Approximately linear.

Most Common Age Quiz

histplot (Age)



35 - 40

Admissions Quiz 1

50%

Admissions Quiz 2

10%

Admissions Quiz 3

80%

Admissions Quiz 4

20%

Gender bias quiz

Yes, in favour of female students.

Aggregation Quiz 1

46%

Aggregation Quiz 2

26%

Gender Bias Revisited Quiz

Yes, in favour of male students.