

STA130H1F - Class #1

Welcome to the Course

Prof. Nathan Taback

2018-10-09

Welcome to STA130

This class

- What is data science?

Welcome to STA130

This class

- What is data science?
- What is statistical reasoning?

Welcome to STA130

This class

- What is data science?
- What is statistical reasoning?
- Introduction to the course (syllabus, website, etc.)

Welcome to STA130

This class

- What is data science?
- What is statistical reasoning?
- Introduction to the course (syllabus, website, etc.)
- Introduction to R and RStudio.

Welcome to STA130

This class



- What is data science?
- What is statistical reasoning?
- Introduction to the course (syllabus, website, etc.)
- Introduction to R and RStudio.
- Distributions of quantitative and categorical variables.

Welcome to STA130





This class

- What is data science?
- What is statistical reasoning?
- Introduction to the course (syllabus, website, etc.)
- Introduction to R and RStudio.
- Distributions of quantitative and categorical variables.
- Plotting distributions using `ggplot2`.








What is data science?

-  +  = data science?











What is data science?

-  +  = data science?
-  +  = data science?











What is data science?

-  +  = data science?
-  +  = data science?
-  +  +  = data science?

What is data science?

-  +  = data science?
-  +  = data science?
-  +  +  = data science?
-  +  +  = data science?

What is data science?

-  +  = data science?
-  +  = data science?
-  +  +  = data science?
-  +  +  = data science?

Data science is an exciting discipline that allows you to turn raw data into understanding, insight, and knowledge. We're going to learn to do this in a tidy way -- more on that later!

Applications of Data Science

- Internet search: Google, Yahoo, Bing, etc. use data science algorithms to rank web pages for a search query.

Applications of Data Science

Recommender Systems: Netflix, Hinge, Amazon, Google, etc. use data science algorithms in recommender systems to suggest products (or dating partners) in accordance with user's interests.

Match Group dating app Hinge to use machine learning for better matches

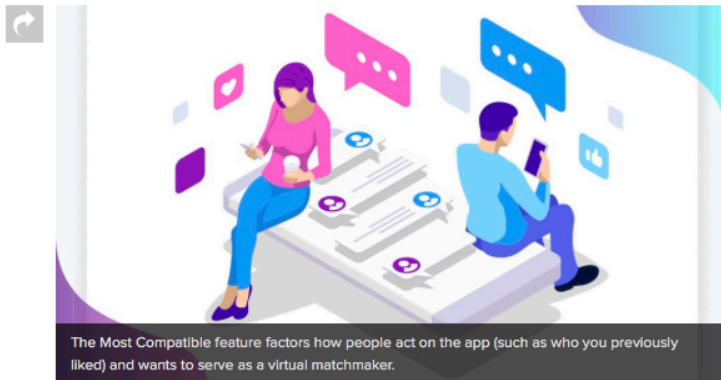
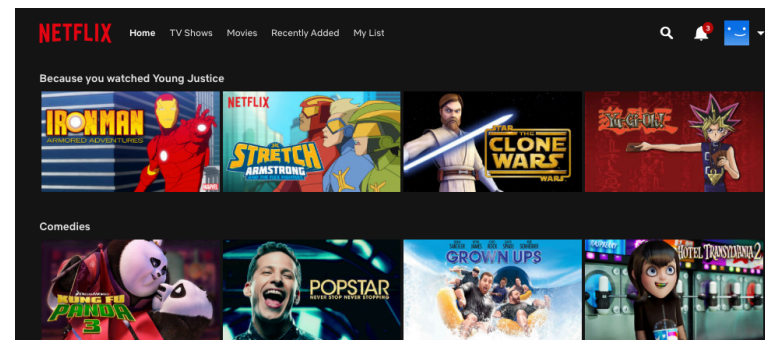


IMAGE: GOLDEN SIKORKA/SHUTTERSTOCK



Applications of Data Science

- Logistics, health care, image and speech recognition, ...

What is statistical reasoning?

- Abraham Wald born in 1902 in Austria.
- Emigrated to the U.S. and eventually became a professor at Columbia.
- During World War II he spent much of his time in the Statistical Research Group (SRG). A classified program that assembled the best American statisticians to the war effort.



What is statistical reasoning?



- The SRG was in an apartment building in NYC a few blocks from Columbia U.

What is statistical reasoning?



- The SRG was in an apartment building in NYC a few blocks from Columbia U.
- The SRG was a very influential group and the military frequently listened to their advice.

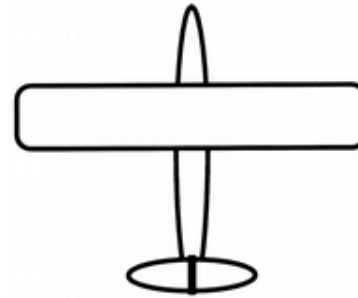
What is statistical reasoning?



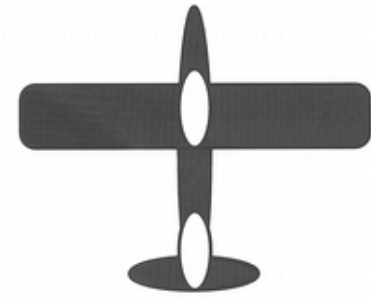
- The SRG was in an apartment building in NYC a few blocks from Columbia U.
- The SRG was a very influential group and the military frequently listened to their advice.
- Wald at the time was still an “enemy alien” , he was not technically allowed to see the reports he was producing.

Missing bullet holes problem

Question: You don't want planes to get shot down by enemy fighters, so you armour them. But armour makes planes heavier, and are less maneuverable and use more fuel. Armouring planes too much is a problem; armouring the planes too little is a problem.



An outline of a plane



A depiction of a plane with shading indicating where returning planes had been shot.

Missing bullet holes problem



Planes were covered in bullet holes, but the holes weren't uniformly distributed across the aircraft.

Missing bullet holes problem

Data from American planes that came back from engagements over Europe.

Question: What parts of the plane has the greatest need for armour?

Section of Plane	Bullet holes per square foot
Engine	1.11
Fuselage (main body of aircraft)	1.73
Fuel system	1.55
Rest of plane	1.8

Missing bullet holes problem

The officers saw an opportunity for efficiency.

Get the same protection with less armour if you concentrate on places with the greatest need.

They asked Wald how much more armour belonged on those parts of the plane.

Section of Plane	Bullet holes per square foot
Engine	1.11
Fuselage (main body of aircraft)	1.73
Fuel system	1.55
Rest of plane	1.8

Missing bullet holes problem

Poll everywhere

Missing bullet holes problem

Wald said that the armour doesn't go where the bullet holes are. It goes where the bullet holes aren't: on the engines.

Missing bullet holes problem

- Wald's insight was to ask: where are the missing holes?

Missing bullet holes problem

- Wald's insight was to ask: where are the missing holes?
- The missing bullet holes were on the missing planes.

Missing bullet holes problem

- Wald's insight was to ask: where are the missing holes?
- The missing bullet holes were on the missing planes.
- The reason planes were coming back with fewer hits to the engine is that planes that got hit in the engine weren't coming back.

Missing bullet holes problem

What did Wald see that the officers who had more knowledge and understanding of aerial combat, couldn't?

Missing bullet holes problem

What did Wald see that the officers who had more knowledge and understanding of aerial combat, couldn't?

- A Statistician is always asking what assumptions are you making? Are they justified?

Missing bullet holes problem

What did Wald see that the officers who had more knowledge and understanding of aerial combat, couldn't?

- A Statistician is always asking what assumptions are you making? Are they justified?
- The officers were making the assumption that the planes that came back were a random sample of all the planes.

Missing bullet holes problem

What did Wald see that the officers who had more knowledge and understanding of aerial combat, couldn't?

- A Statistician is always asking what assumptions are you making? Are they justified?
- The officers were making the assumption that the planes that came back were a random sample of all the planes.
- Once you recognize that you have been making this hypothesis, it takes a moment to realize that it's wrong.

Missing bullet holes problem

What did Wald see that the officers who had more knowledge and understanding of aerial combat, couldn't?

- A Statistician is always asking what assumptions are you making? Are they justified?
- The officers were making the assumption that the planes that came back were a random sample of all the planes.
- Once you recognize that you have been making this hypothesis, it takes a moment to realize that it's wrong.
- In statistical lingo, the rate of survival and location of bullet holes are correlated.

Survivorship Bias

- The underlying statistical phenomena is often called survivorship bias.

Survivorship Bias

- The underlying statistical phenomena is often called survivorship bias.
- Thinking statistically let's you see the common skeleton shared by problems that look very different on the surface.

Survivorship Bias

- The underlying statistical phenomena is often called survivorship bias.
- Thinking statistically lets you see the common skeleton shared by problems that look very different on the surface.
- Thus you have meaningful experience even in areas where you appear to have none.

Who am I?

✉ nathan.taback@utoronto.ca

🏠 <http://sta130.utstat.toronto.edu>

🏛️ Sidney Smith, SS6027C

📅 Monday 12:30 - 14:00 (after class I'll go to my office).

What is this course?

Everything you want to know about the course, and everything you will need for the course will be posted at

<http://sta130.utstat.toronto.edu>

What is this course?

Everything you want to know about the course, and everything you will need for the course will be posted at

<http://sta130.utstat.toronto.edu>

- Will we be doing computing? Yes.

What is this course?

Everything you want to know about the course, and everything you will need for the course will be posted at

<http://sta130.utstat.toronto.edu>

- Will we be doing computing? Yes.
- Is this an intro CS course? No, but many themes are shared.

What is this course?

Everything you want to know about the course, and everything you will need for the course will be posted at

<http://sta130.utstat.toronto.edu>

- Will we be doing computing? Yes.
- Is this an intro CS course? No, but many themes are shared.
- Is this an intro stat course? Yes, but it's not your high school statistics course.

What is this course?

Everything you want to know about the course, and everything you will need for the course will be posted at

<http://sta130.utstat.toronto.edu>

- Will we be doing computing? Yes.
- Is this an intro CS course? No, but many themes are shared.
- Is this an intro stat course? Yes, but it's not your high school statistics course.
- What computing language will we learn? R.

What is this course?

Everything you want to know about the course, and everything you will need for the course will be posted at

<http://sta130.utstat.toronto.edu>

- Will we be doing computing? Yes.
- Is this an intro CS course? No, but many themes are shared.
- Is this an intro stat course? Yes, but it's not your high school statistics course.
- What computing language will we learn? R.
- Why not language X? We can discuss that over ☕.

Create an RStudio.cloud account

- List steps

World happiness

What influences your happiness?

We'll look at the data from the [World Happiness Report 2017](#)

[Video](#)

Data from the Gallup World Poll is in the file `happinessdata_2017.csv`.

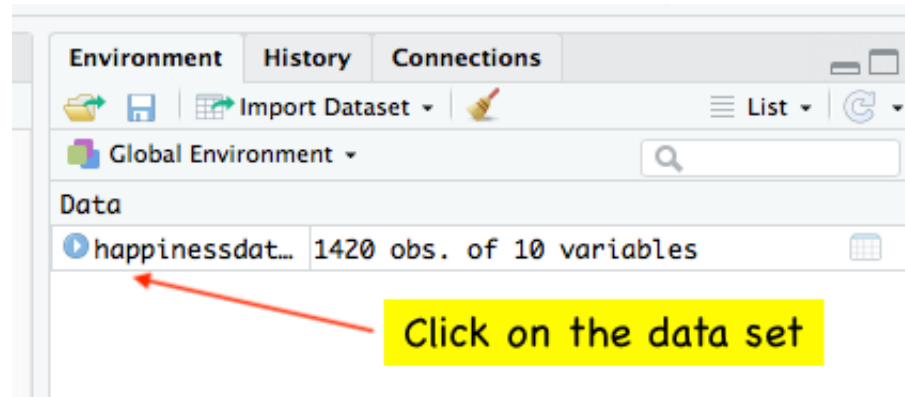
Read the data into R

```
# Read in the data  
happinessdata_2017 <- read_csv("happinessdata_2017.csv")
```

View the data

There are two ways to view a data set in RStudio.

1. Click on the Environment tab in the upper right hand corner (Environment, History, Connections pane). Then click on the data set



1. Type `glimpse(happinessdata_2017)` in an R code chunk.

```
# type the command here
```

Questions:

1. What does each row and column represent?
2. How many rows and columns are in happinessdata_2017?

How is happiness measured and ranked?

The rankings are based on answers to the main life evaluation question asked in the poll. This is called the Cantril ladder: it asks respondents to think of a ladder, with the best possible life for them being a 10, and the worst possible life being a 0. They are then asked to rate their own current lives on that 0 to 10 scale. (Ref: <http://worldhappiness.report/faq/>)

What is the distribution of the Cantril ladder?

- We need to figure out the variable name of Cantril ladder.
- `life_ladder` — average response for a country to the question:
“Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?”
- The `life_ladder` variable is an example of a **numerical (quantitative) variable**. A quantitative variable takes numerical values that are ordered and differences are meaningful.
- The **distribution** of a variable tells us what values it takes and how often it takes these values.

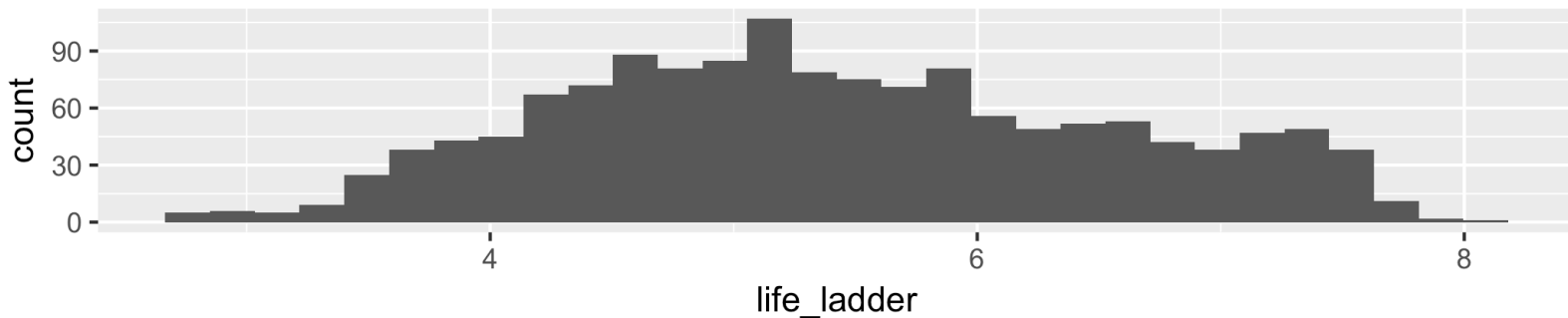
Examining the distribution of life_ladder: histogram

This code

```
library(tidyverse)
ggplot(data = happinessdata_2017) +
  aes(x = life_ladder) +
  geom_histogram()
```

produces this plot (histogram)

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Graphical exploration of data

We'll use the `ggplot2` package in R to construct our graphs.

“gg” = Grammar of Graphics (Leland Wilkinson), a structure to combine graphical elements together to make a meaningful display of data

To use `ggplot2` functions, need to first load the package `ggplot2`, which is also part of the `tidyverse` package.

```
library(tidyverse)
```

ggplot2

In ggplot2, the structure of the code to produce most plots is

```
ggplot(data=[dataset],  
       aes(x=[var1],  
          y=[var2])) +  
geom_xxx( ) +  
other options
```

- **aesthetic**: mapping between a variable and where it will be represented on the graph (e.g., x axis, colour-coding, etc.)
- **geometry**: what are you plotting (e.g., points , lines, histogram, etc.)
 - Every plot must have at least one geometry and there is no upper limit
 - You add a geometry to a plot using +

histogram using ggplot2

```
library(tidyverse)
ggplot(data = happinessdata_2017) +
  aes(x = life_ladder) +
  geom_histogram()
```

- Just need one aesthetic, x.

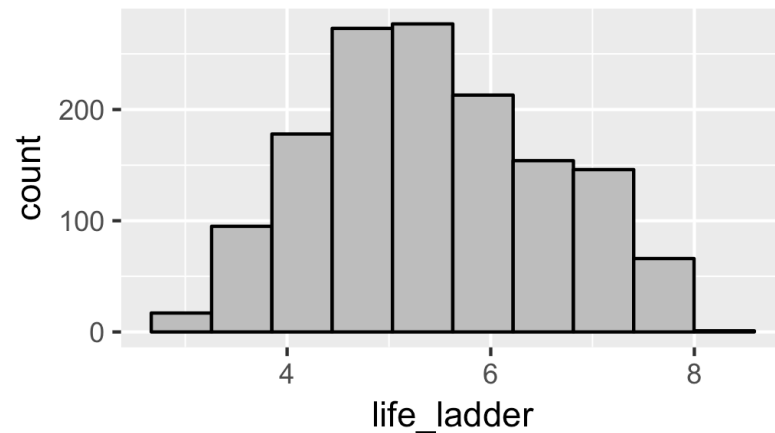
Constructing a histogram

- Count the number of numerical values that lie within ranges, called bins.
- Bins are defined by their lower bounds (inclusive); the upper bound is the lower bound of the next bin.
- Histogram displays the distribution (count (default) or density) of the numerical values in the bins.
- Horizontal axis is numerical, hence no gaps.

Number of Bins of a histogram

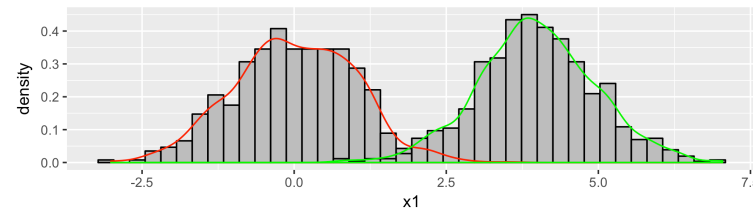
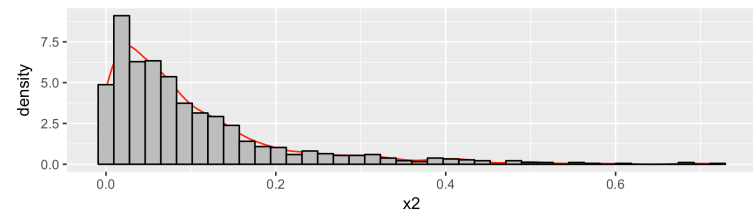
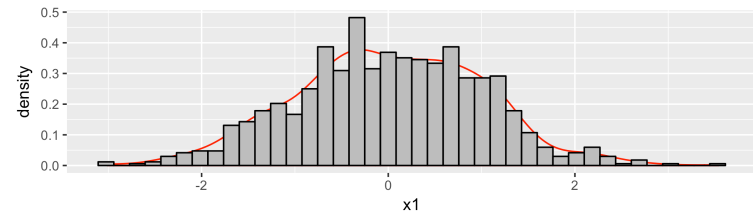
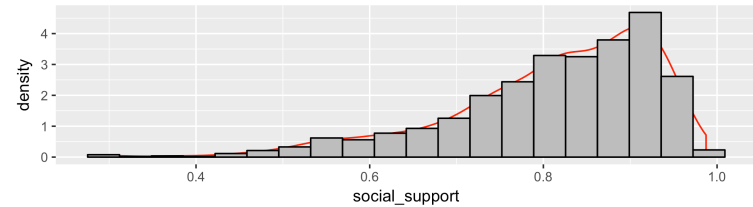
```
library(tidyverse)
ggplot(data = happinessdata_2017) +
  aes(x = life_ladder) +
  # colour is outline of bin
  geom_histogram(bins = 2, colour = "black", fill = "grey")

ggplot(data = happinessdata_2017) +
  aes(x = life_ladder) +
  geom_histogram(bins = 10, colour = "black", fill = "grey")
```



Properties of distributions

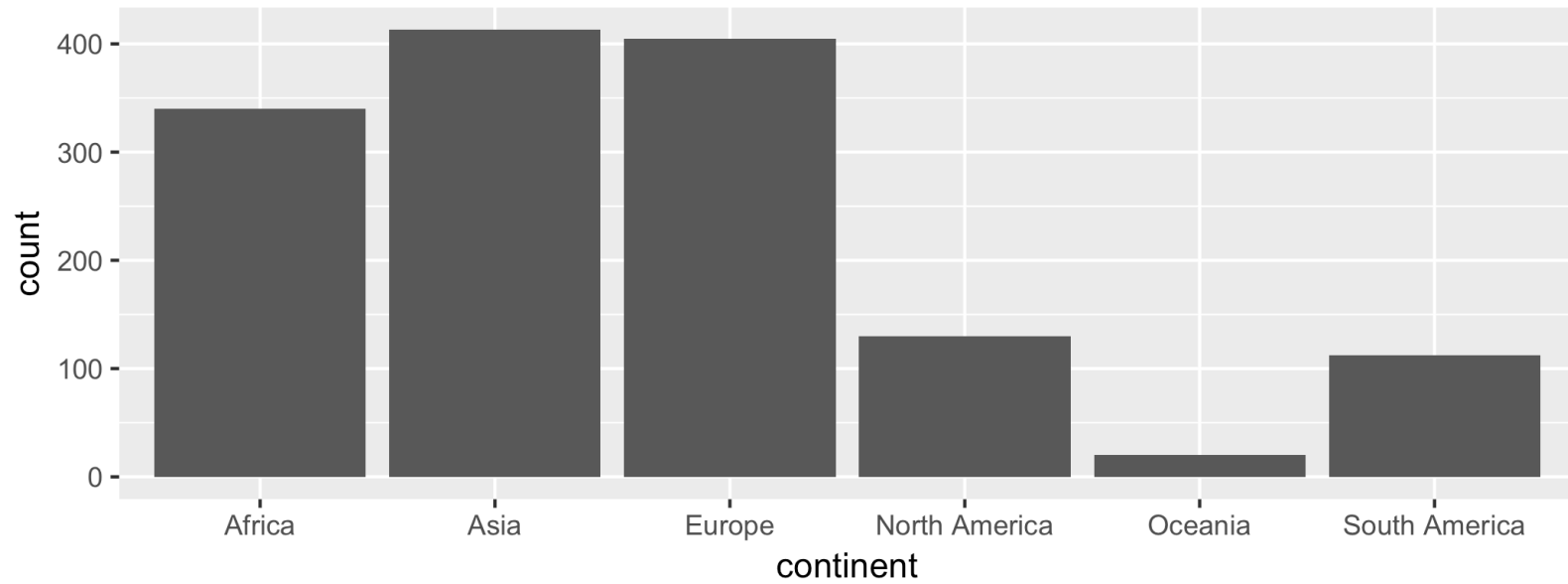
- **Shape** of the distribution:
 - could be *symmetric*, *left-skewed*, *right-skewed* (skew is to the direction of the longer tail)
 - number of **modes (peaks)**: unimodal, bimodal, multimodal, uniform
 - unusual observations



Basic plots for a categorical variable: bar plot

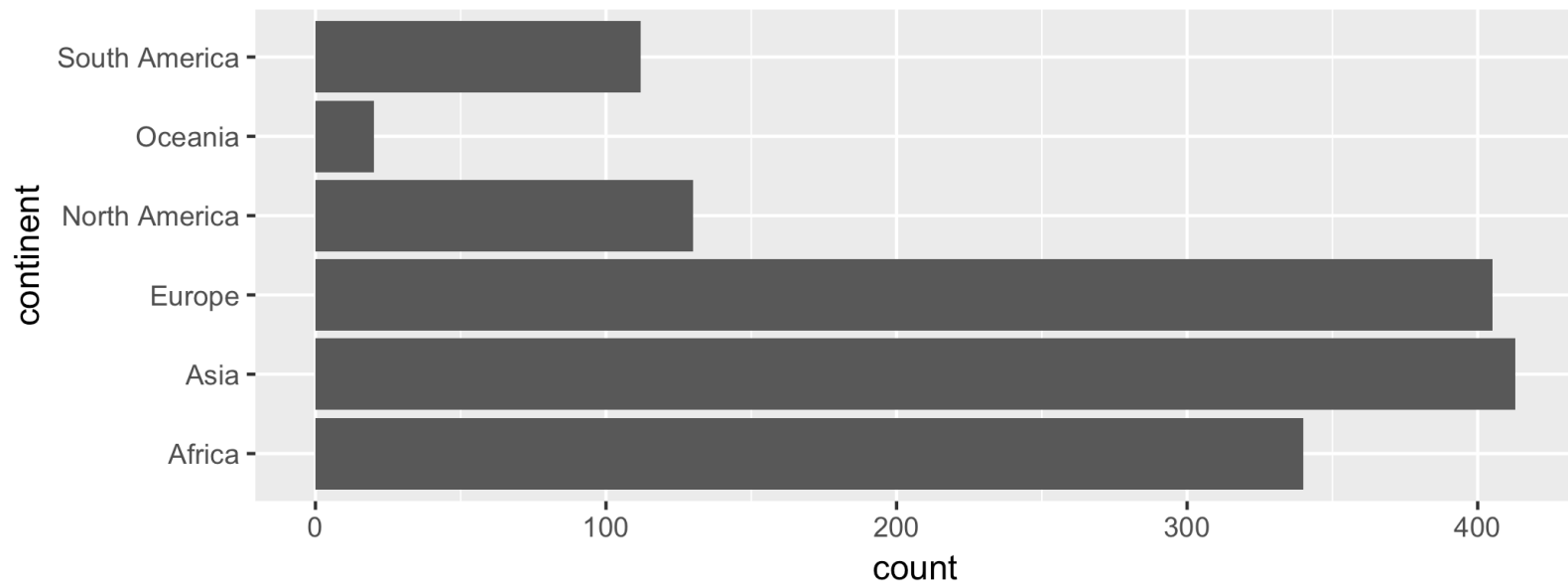
- Displays the distribution of a categorical variable, the frequency of its different values
- Heights (or lengths) of bars are proportional to the percent of individuals
- Bars have arbitrary (but equal) widths and spacings

```
ggplot(data = happinessdata_2017, aes(x = continent)) +  
  geom_bar()
```



An alternative, particularly useful for long labels

```
ggplot(data = happinessdata_2017, aes(x = continent)) +  
  geom_bar() +  
  coord_flip()
```

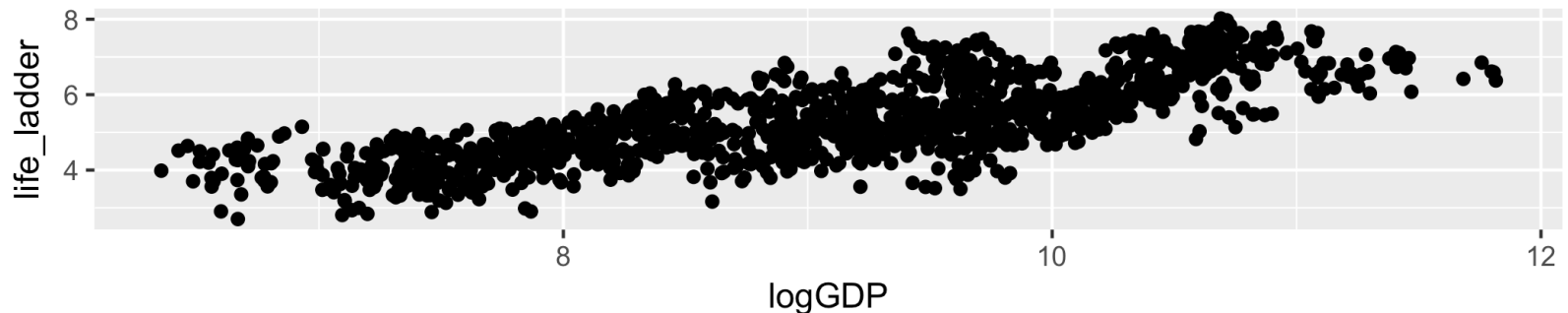


Looking at the relationship between two variables

What is the relationship between happiness and wealth?

```
library(tidyverse)
ggplot(data = happinessdata_2017) +
  aes(x = logGDP, y = life_ladder) +
  geom_point()
```

```
## Warning: Removed 35 rows containing missing values (geom_point).
```

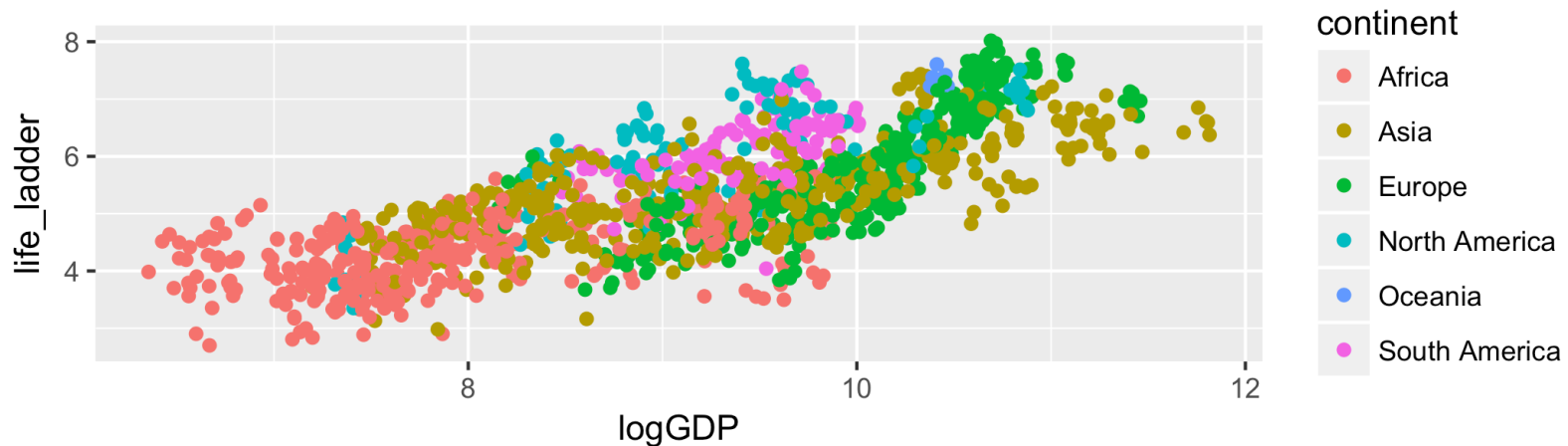


A **scatterplot** of `life_ladder` versus `logGDP` consists of points representing a countries with values of both `life_ladder` and `logGDP`. What happens if one of the values is missing for a country?

What is the relationship between happiness, wealth, and continent?

```
library(tidyverse)
ggplot(data = happinessdata_2017, ) +
  aes(x = logGDP, y = life_ladder, colour = continent) +
  geom_point()
```

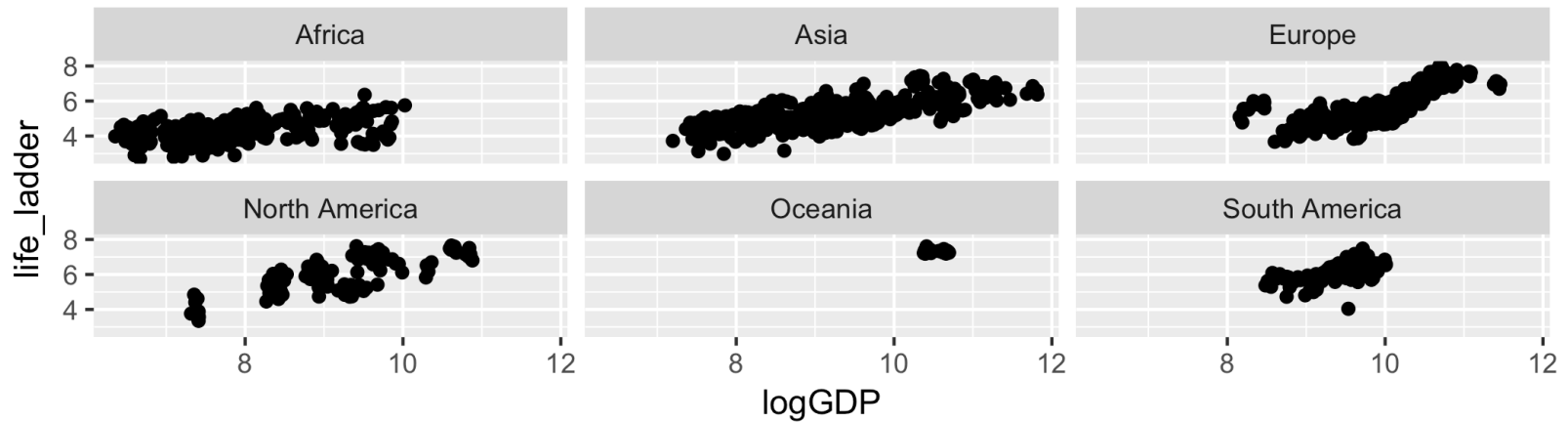
```
## Warning: Removed 35 rows containing missing values (geom_point).
```



Colour the points by continent.

```
library(tidyverse)
ggplot(data = happinessdata_2017, ) +
  aes(x = logGDP, y = life_ladder) +
  geom_point() +
  facet_wrap(~continent)
```

```
## Warning: Removed 35 rows containing missing values (geom_point).
```



`facet_wrap()` produces a sequence of rectangular plots.