

# **STAI30H1F(L201) - Class #1**

## **Welcome to the Course**

Prof. Nathalie Moon











2018-10-09

# Welcome to STAI30

## This class

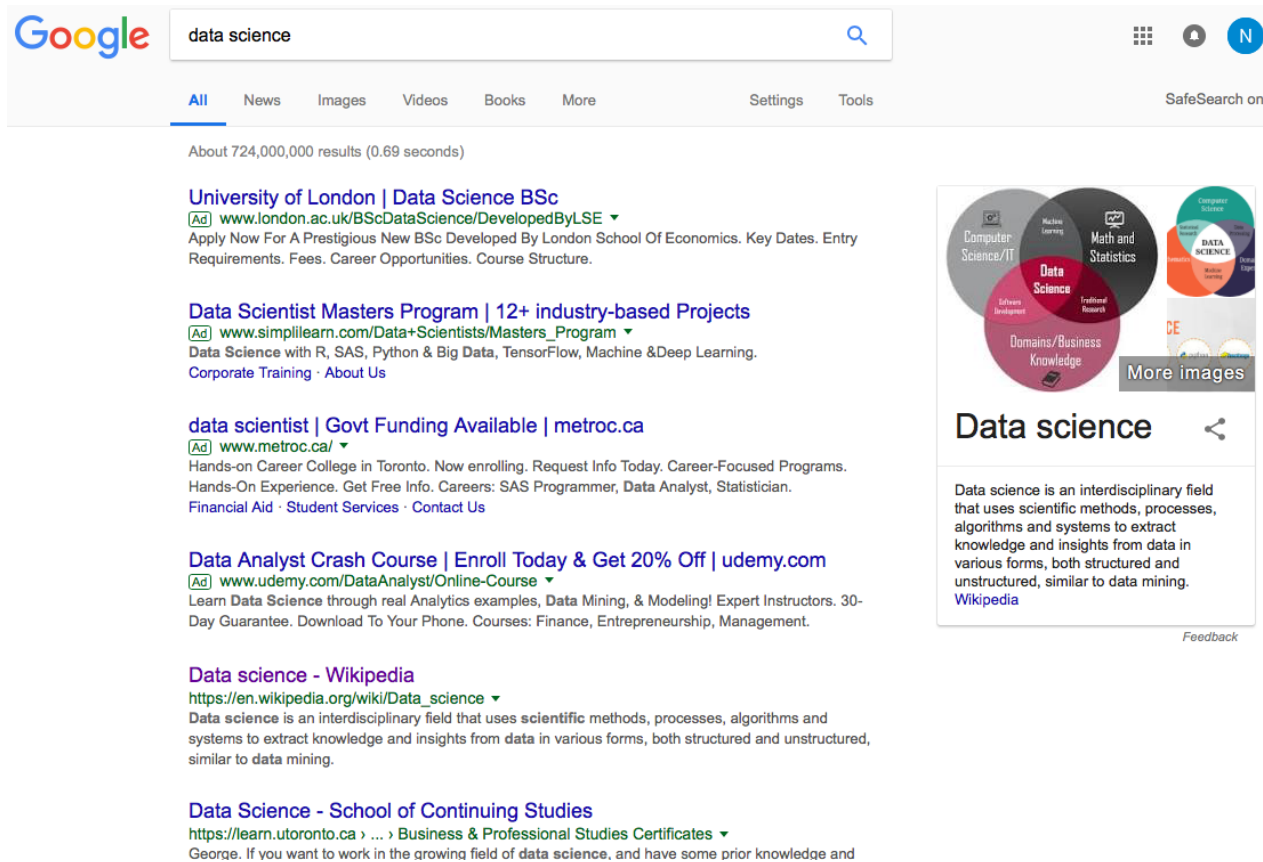
- What is data science?
- What is statistical reasoning?
- Introduction to the course (syllabus, website, etc.)
- Introduction to R and RStudio.
- Distributions of quantitative and categorical variables.
- Plotting distributions using `ggplot2`.

# What is data science?

-  +  = data science?
  -  +  = 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.



The image shows a Google search interface for the query "data science". The search bar at the top left contains the text "data science" and a magnifying glass icon. Below the search bar are navigation tabs for "All", "News", "Images", "Videos", "Books", and "More", along with "Settings" and "Tools". The search results are displayed below, starting with "About 724,000,000 results (0.69 seconds)".

The search results include several entries:

- University of London | Data Science BSc**  
An advertisement for a BSc program developed by LSE, with a link to [www.london.ac.uk/BScDataScience/DevelopedByLSE](http://www.london.ac.uk/BScDataScience/DevelopedByLSE). The description mentions applying for a prestigious new BSc and lists details like key dates, entry requirements, fees, and career opportunities.
- Data Scientist Masters Program | 12+ industry-based Projects**  
An advertisement for a masters program on [www.simplilearn.com/Data+Scientists/Masters\\_Program](http://www.simplilearn.com/Data+Scientists/Masters_Program). It highlights hands-on experience with R, SAS, Python, and Big Data, as well as TensorFlow, Machine & Deep Learning, and corporate training options.
- data scientist | Govt Funding Available | metro.c.ca**  
An advertisement for a career college in Toronto ([www.metro.ca/](http://www.metro.ca/)) offering hands-on experience and career-focused programs. It lists careers like SAS Programmer, Data Analyst, and Statistician, and provides links for financial aid, student services, and contact.
- Data Analyst Crash Course | Enroll Today & Get 20% Off | udemy.com**  
An advertisement for a crash course on [www.udemy.com/DataAnalyst/Online-Course](http://www.udemy.com/DataAnalyst/Online-Course). It promises to teach data science through real analytics examples, data mining, and modeling, with expert instructors and a 30-day guarantee.
- Data science - Wikipedia**  
A link to the Wikipedia page for data science ([https://en.wikipedia.org/wiki/Data\\_science](https://en.wikipedia.org/wiki/Data_science)). The snippet defines data science as an interdisciplinary field using scientific methods to extract knowledge from data.
- Data Science - School of Continuing Studies**  
A link to a program at the University of Toronto (<https://learn.utoronto.ca>) for business and professional studies. It mentions George and notes that the program is for those with prior knowledge in the field.

On the right side of the search results, there is a knowledge panel for "Data science". It features a circular diagram with overlapping circles labeled "Computer Science/IT", "Math and Statistics", "Data Science", "Software Development", "Machine Learning", "Big Data", and "Domains/Business Knowledge". Below the diagram is a "More images" button. The panel also contains a definition of data science and a link to the Wikipedia page. A "Feedback" link is located at the bottom right of the panel.

# 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

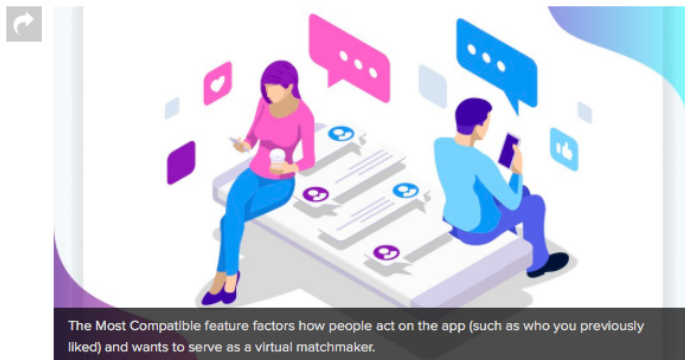
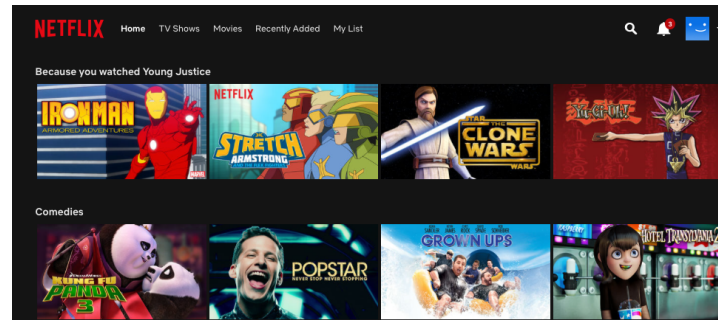


IMAGE: GOLDEN SIKORKA/SHUTTERSTOCK



# 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.

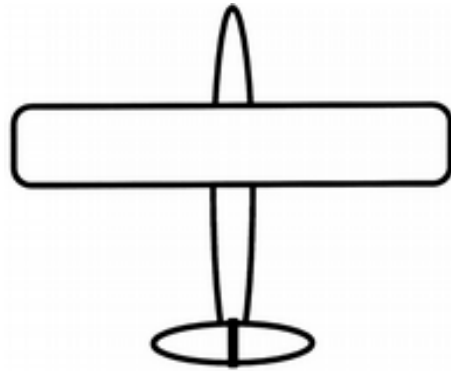
# What is statistical reasoning?



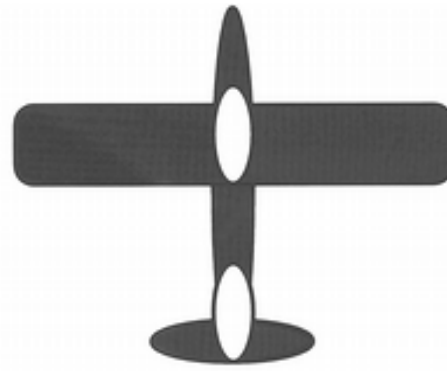
- The SRG was in an apartment building in NYC a few blocks from Columbia U.
- The SRG was a very influential group and the military frequently listened to their advice.
- Wald at the time was still an “enemy alien”, he was not technically allowed to see the reports he was producing.



# Missing bullet holes problem



An outline of a plane



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

**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.

# 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: Which parts of the plane have 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




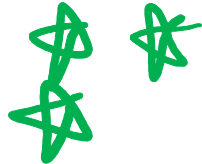
# Missing bullet holes problem

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



What do you think?

Go to [pollev.com/loop](https://pollev.com/loop)

# 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.

- strategic aiming?

\*- planes that are hit in more critical areas (engine!) crash in the field and not return to the base

↳ not observable

# 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 ones that crashed)
- 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?

- 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.
- 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.

↳ survivorship bias also arises in other settings such as medical studies.



# Who am I?

-  [nathalie.moon@utoronto.ca](mailto:nathalie.moon@utoronto.ca)
-  <http://stat130.utstat.toronto.edu>
-  Sidney Smith, SS6024A
-  Monday 4:00-5:30 (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>

- 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

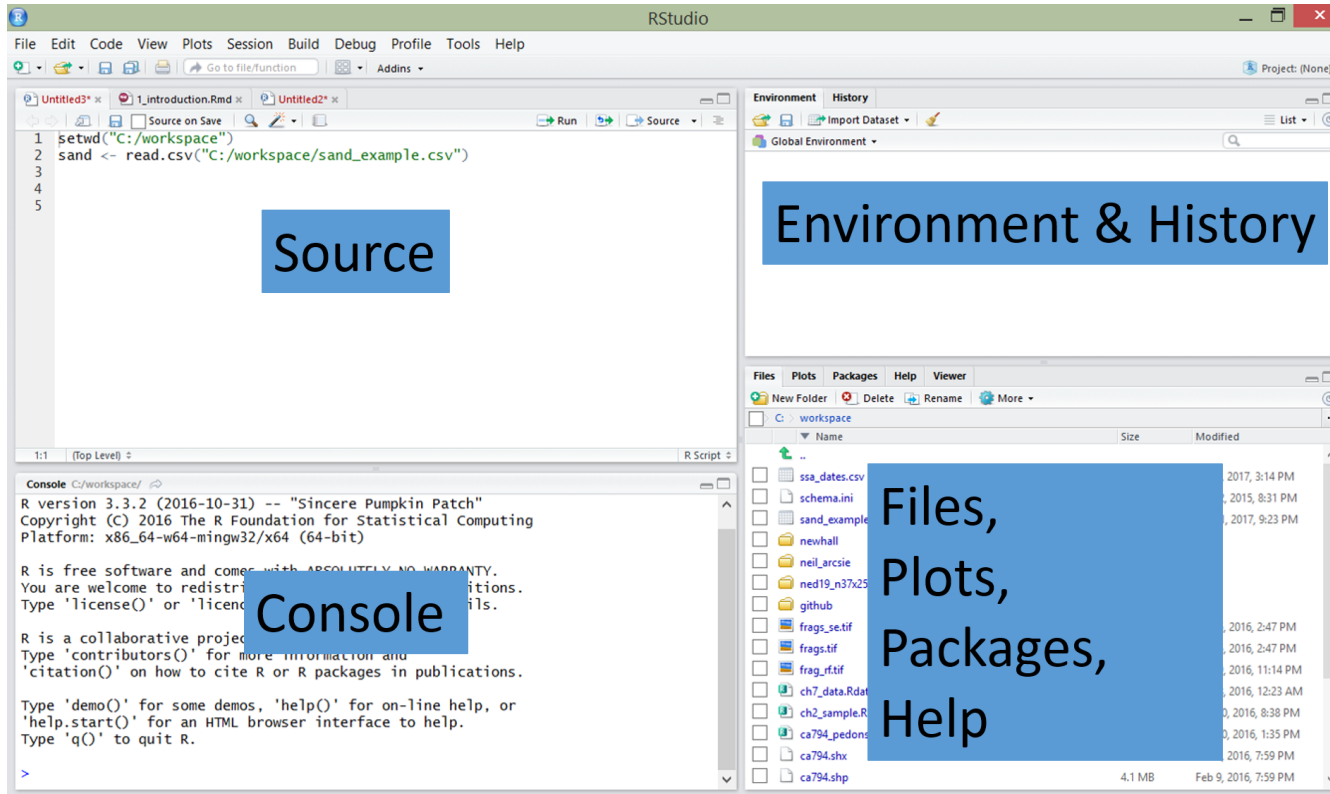
1. Go to [rstudio.cloud](https://rstudio.cloud) and click on “Get Started”
2. Create an account
3. Click [here](#) to join the workspace for STA130

When you log in to Rstudio.cloud, you’ll now have two “Spaces” on the left hand side:

- YourWorkspace
- STA130\_Fall2018

Instructions for getting started with Week 1 Practice Problems can be found on the [course webpage](#)

# Introduction to RStudio



- Create a project
- Create a notebook (.Rmd file)
- Insert a code chunk

# 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



2. 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?

rows = observations → 1420 rows  
columns = variables → 10 columns



# 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's the distribution of the Cantril ladder?

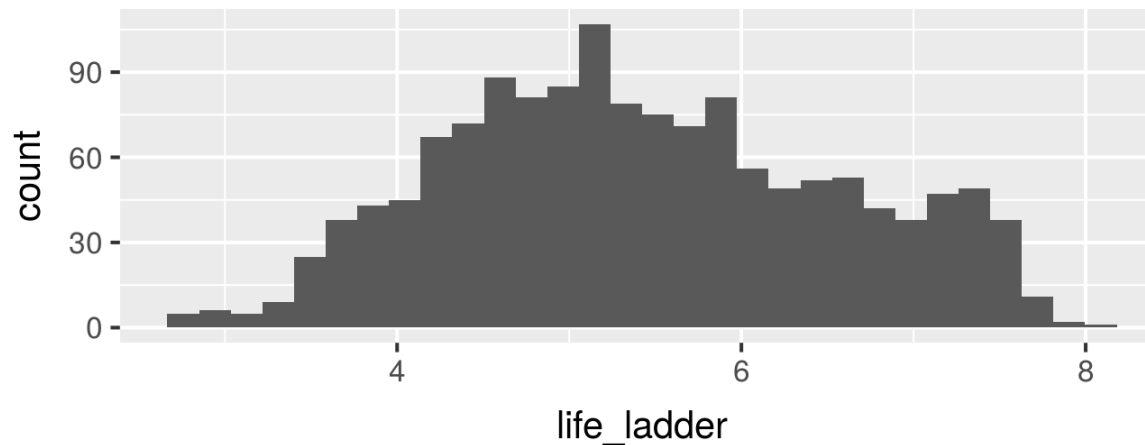
- We need to figure out the variable name of Cantril ladder.
- `life_ladder` — average response for a country to the question:  
*“Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the 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

```
library(tidyverse)
ggplot(data = happinessdata_2017) +
  aes(x = life_ladder) +
  geom_histogram() +
  theme_gray(base_size=5)
```

produces this plot (histogram)

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



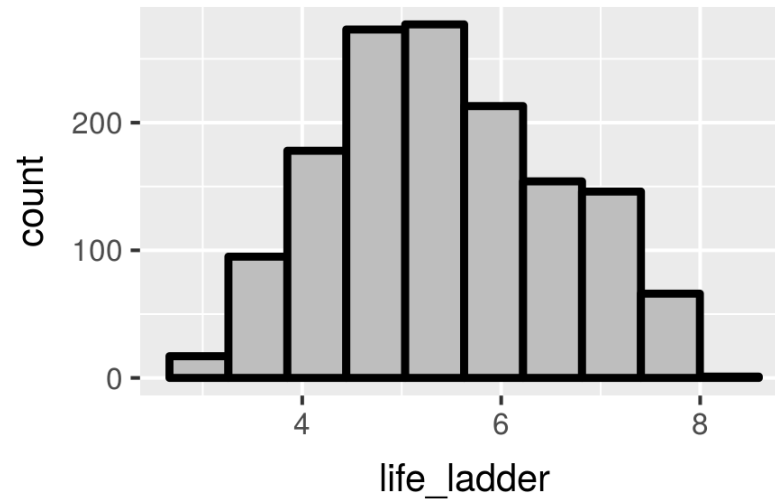
# Constructing a histogram

- 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

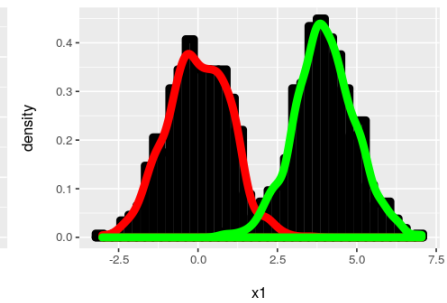
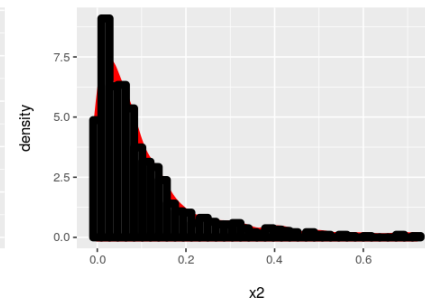
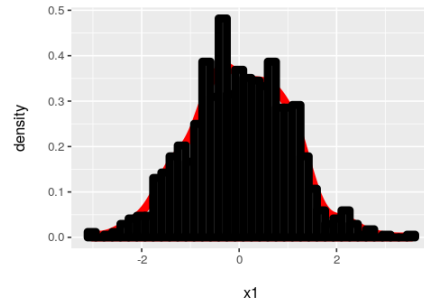
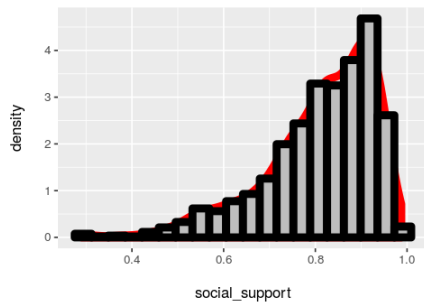
```
library(tidyverse)
ggplot(data = happinessdata_2017) +
  aes(x = life_ladder) +
  # colour is outline of bin
  geom_histogram(bins = 3, colour = "black", fill = "grey")

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



# 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.)

Notes:

- 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.
- Why?

↳ because only plotting one variable

# Basic plot for a categorical variable: bar plot

## Categorical variable:

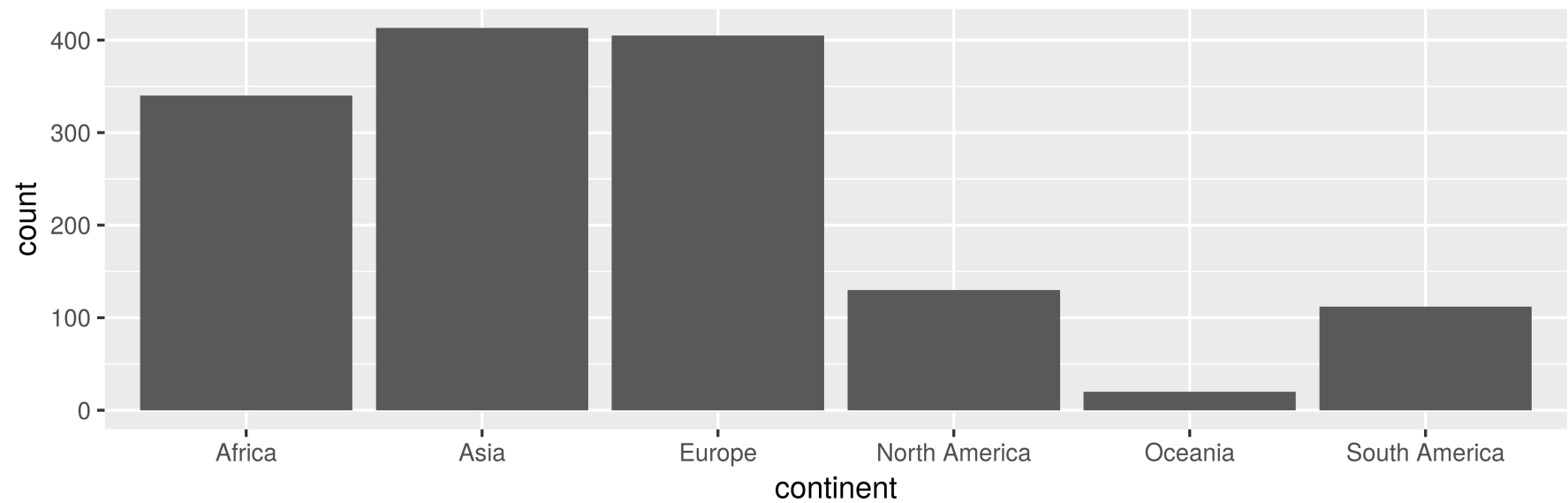
- A categorical variable takes a discrete number of values that are not ordered (e.g. country, continent, etc)
- Sometimes these may be coded as numbers in the data (e.g. male = 1, female = 0), but there is no meaningful ordering and the numerical differences are not important.

## 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

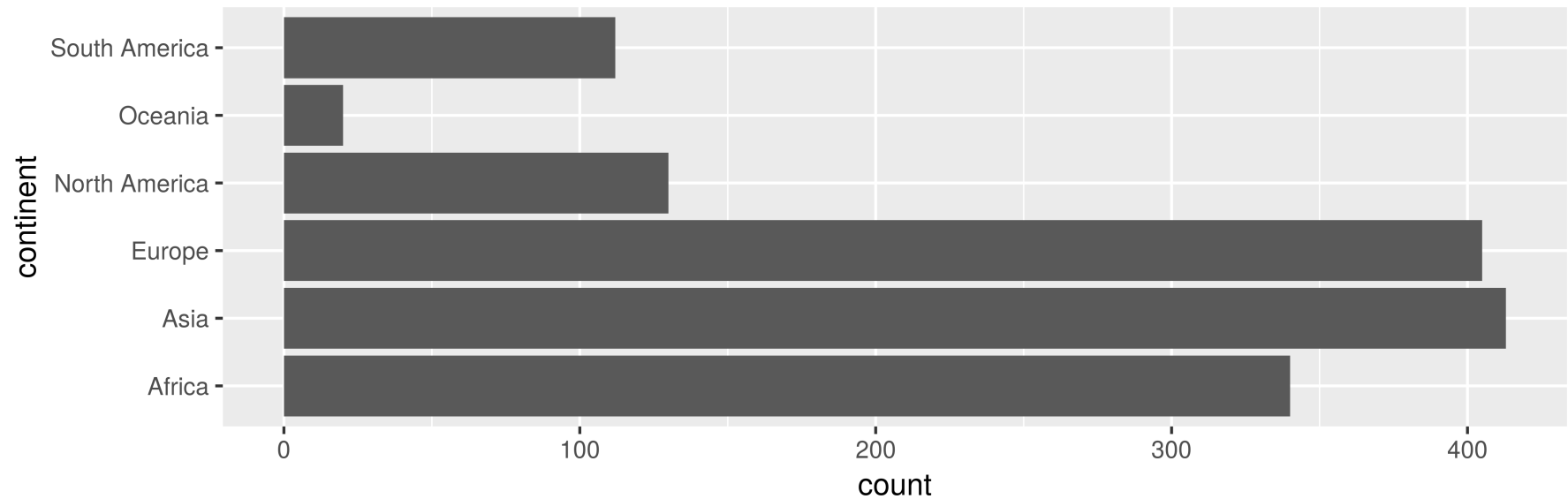
# Bar plot example

```
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()
```

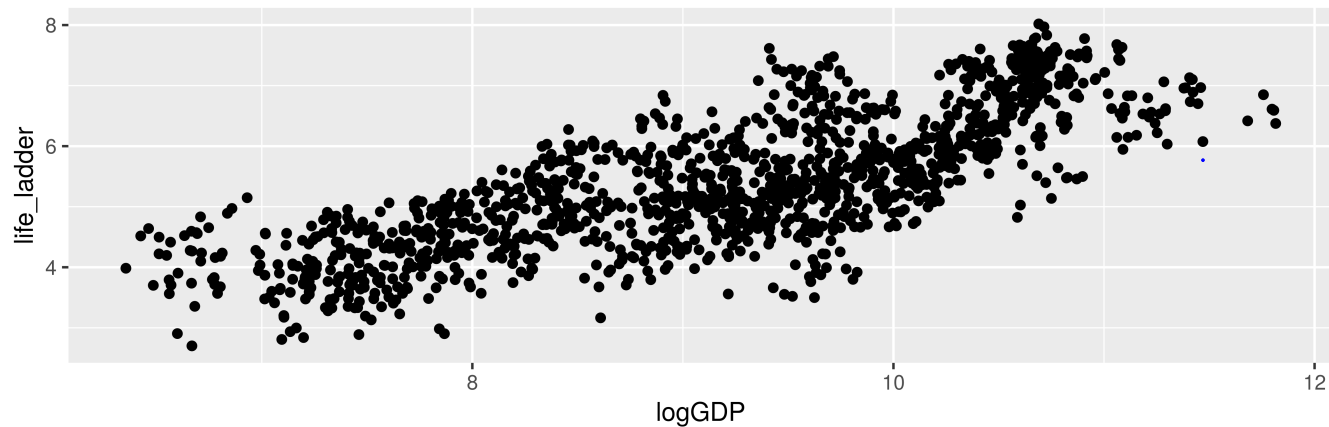


# Looking at the relationship between two variables

What is the relationship between happiness and wealth?

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

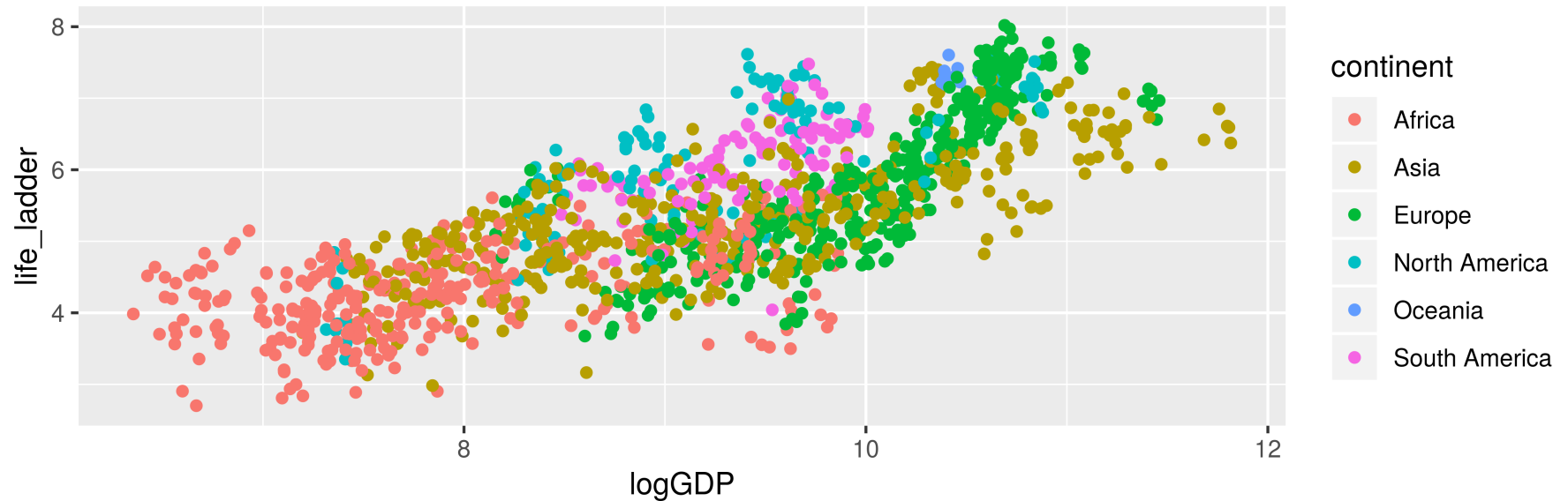
```
## Warning: Removed 35 rows containing missing values (geom_point).
```



# 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()
```

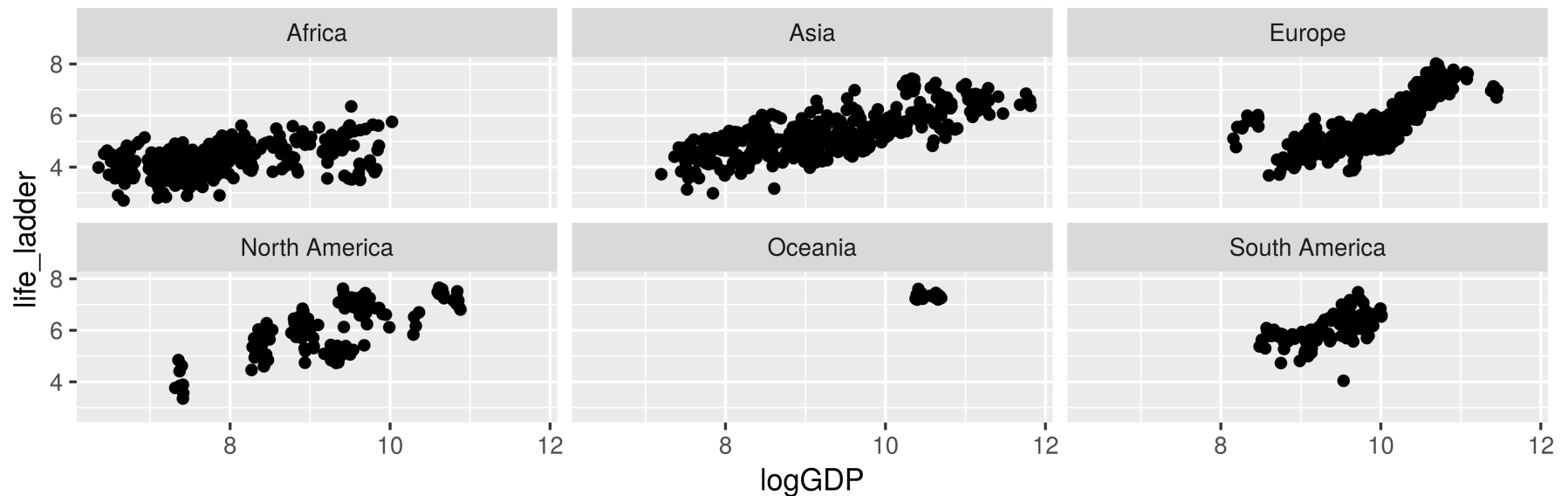
```
## Warning: Removed 35 rows containing missing values (geom_point).
```



Colour the points by continent.

# Alternate way of looking at the relationship between three variables

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



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

# To do this week

- Create an account with RStudio.cloud
- Join the [workspace](#) for STA130
- Work on the week 1 practice problems and bring your solutions to your tutorial on Friday