Visualization_I_Class4-Completed

February 18, 2021

[1]:
```
# import necessary packages
import pandas as pd
import numpy as np
from plotnine import *

import warnings
warnings.filterwarnings('ignore')
```

1. Load the necessary libraries (pandas, plotnine)
2. Load the following dataset `Fifa = pd.read_csv("https://raw.githubusercontent.com/cmparlettpelle` For more info check out this link.
3. How old is the average player in FIFA20? Use plotnine/ggplot to plot a histogram of their ages.
4. Do right footed players weigh more than right footed players? Use ggplot/plotnine to make a graph to answer this question.
5. Is there a difference in height between the clubs Chelsea and Manchester United? Use ggplot/plotnine to make a graph to answer this question.
6. Is there a difference in the relationship between height and weight for people born in each of the 12 months? Use ggplot/plotnine to make a graph to answer this question. (see code below on how to extract the month from the column `fifa[dob]`)
7. Is there an averaged difference in weight between players with different `body_types`? Use ggplot/plotnine to make a graph to answer this question. Discuss what kind of plot would best communicate this information in your opinion.
8. Calculate a new column `bmi` for the fifa dataset. The formula for BMI is below. Note: Body Mass Index (BMI) is not a universal metric of health, and should NOT be taken as such. Use ggplot/plotnine to plot a histogram of the different BMIs in the dataset. Add a dashed line using `+ geom_vline(xintercept = mean, linetype = "dashed", size = 3)` where mean is the mean BMI for the whole dataset.

$$BMI = \frac{weight(kg)}{height(m)^2}$$

9. What is the relationship between height and weight for each `body_type`? Use `facet_wrap()` to make a separate height/weight scatterplot for each body type.

10. Is there an association between jersey number and age? Use ggplot/plotnine to make a graph to answer this question.

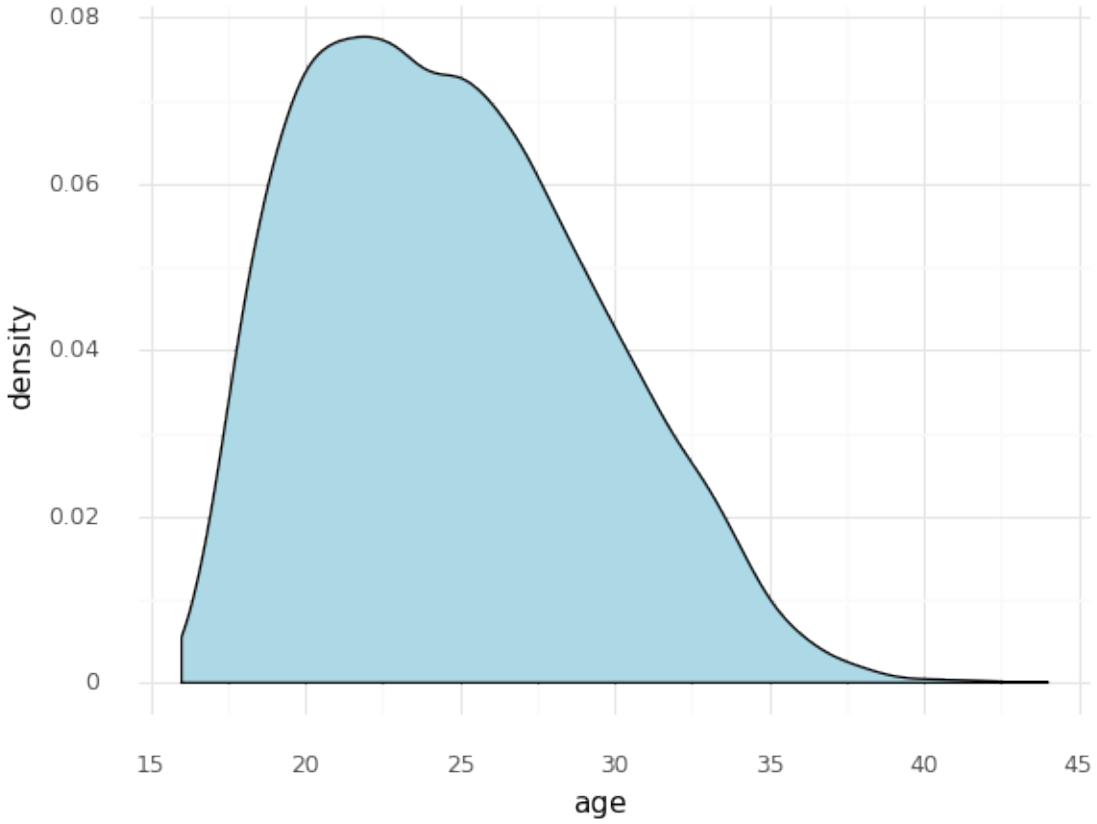
11. Let's use your CPSC230 skills, create a column in `fifa` called `name_len` that counts the number of characters in each player's `long_name` (spaces shouldn't count). Then use ggplot/plotnine to create a histogram of `name_len` and add `+theme_minimal()` to your graph. What is the typical range of name length?

[3]:
```
### YOUR CODE HERE ###
fifa = pd.read_csv("https://raw.githubusercontent.com/cmparlettpelleriti/
↪CPSC392ParlettPelleriti/master/Data/players_15.csv")

# get month of date of birth
fifa["monthBorn"] = fifa["dob"].str.extract(r'-([0-9][0-9])-')

fifa.head()
```

[3]:

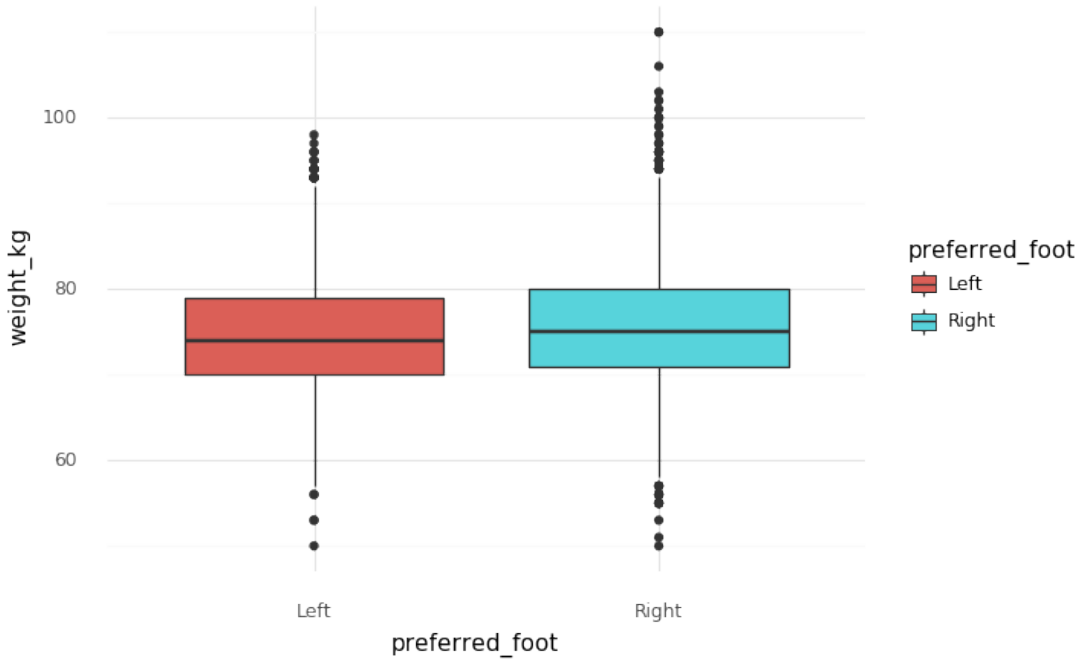
```
   sofifa_id                                        player_url  \
0     158023  https://sofifa.com/player/158023/lionel-messi/…
1      20801  https://sofifa.com/player/20801/c-ronaldo-dos-…
2       9014  https://sofifa.com/player/9014/arjen-robben/15…
3      41236  https://sofifa.com/player/41236/zlatan-ibrahim…
4     167495  https://sofifa.com/player/167495/manuel-neuer/…

           short_name                             long_name  age         dob  \
0            L. Messi       Lionel Andrés Messi Cuccittini   27  1987-06-24
1  Cristiano Ronaldo  Cristiano Ronaldo dos Santos Aveiro   29  1985-02-05
2          A. Robben                          Arjen Robben   30  1984-01-23
3     Z. Ibrahimović                  Zlatan Ibrahimović   32  1981-10-03
4            M. Neuer                         Manuel Neuer   28  1986-03-27

   height_cm  weight_kg  nationality                club  …   ldm   cdm  \
0        169         67    Argentina         FC Barcelona  …  62+3  62+3
1        185         80     Portugal          Real Madrid  …  63+3  63+3
2        180         80  Netherlands    FC Bayern München  …  64+3  64+3
3        195         95       Sweden  Paris Saint-Germain  …  65+3  65+3
4        193         92      Germany    FC Bayern München  …   NaN   NaN

    rdm   rwb    lb   lcb    cb   rcb    rb monthBorn
0  62+3  62+3  54+3  45+3  45+3  45+3  54+3        06
1  63+3  63+3  57+3  52+3  52+3  52+3  57+3        02
2  64+3  64+3  55+3  46+3  46+3  46+3  55+3        01
3  65+3  61+3  56+3  55+3  55+3  55+3  56+3        10
4   NaN   NaN   NaN   NaN   NaN   NaN   NaN        03

[5 rows x 105 columns]
```

[7]:
```
#3
ggplot(fifa, aes("age")) + geom_histogram(fill = "lightblue", color = "black")␣
↪+ theme_minimal()

#or

ggplot(fifa, aes("age")) + geom_density(fill = "lightblue", color = "black") +␣
↪theme_minimal()
```



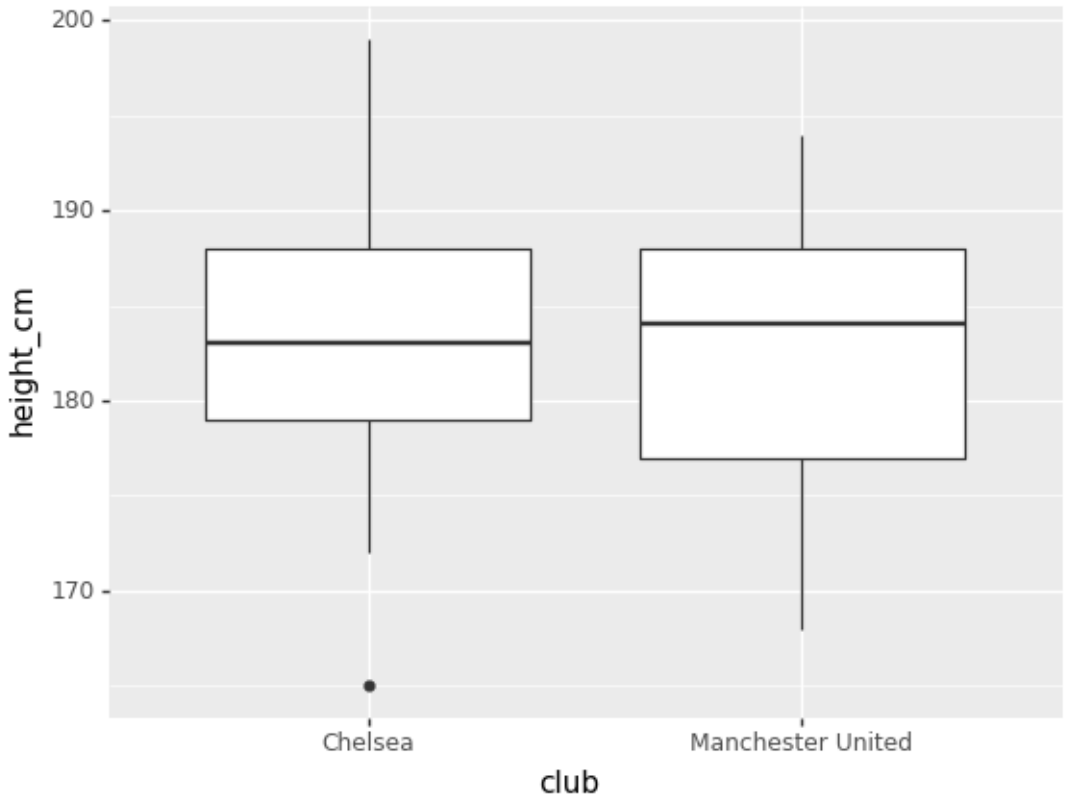
[7]:
```
<ggplot: (8773510942874)>
```

[9]:
```
#4
ggplot(fifa, aes(x = "preferred_foot", y = "weight_kg", fill =␣
↪"preferred_foot")) + geom_boxplot() + theme_minimal()
```



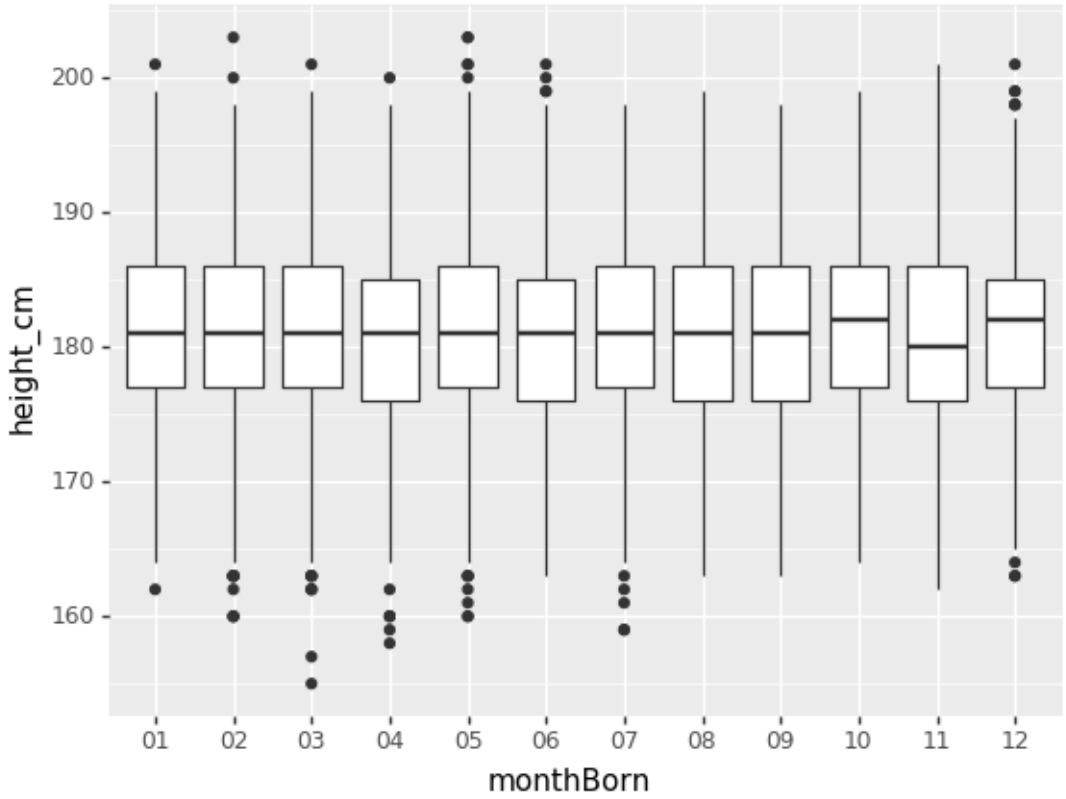
[9]: <ggplot: (8773510943404)>

[26]:
```
#5

CorManU = (fifa["club"] == "Chelsea") | (fifa["club"] == "Manchester United")

CandManU = fifa.loc[CorManU]

ggplot(CandManU, aes(x = "club", y = "height_cm")) + geom_boxplot()
```



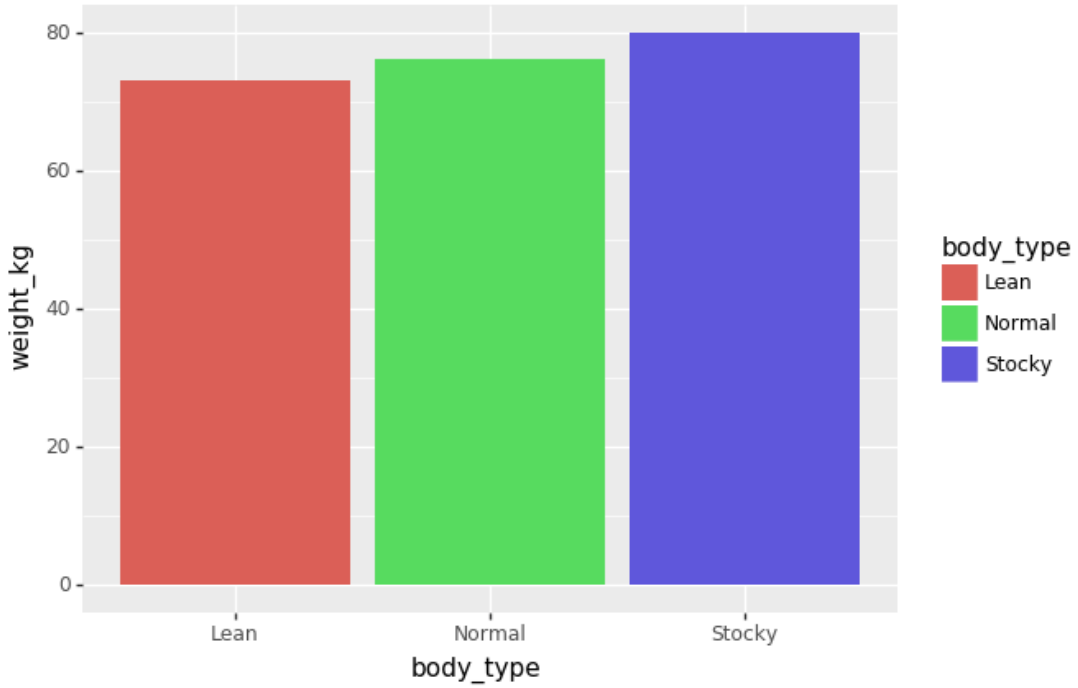
[26]: <ggplot: (8773494999239)>

[27]:
```
#6

ggplot(fifa, aes(x = "monthBorn", y = "height_cm", fill = "height_cm")) +␣
↪geom_boxplot()
```



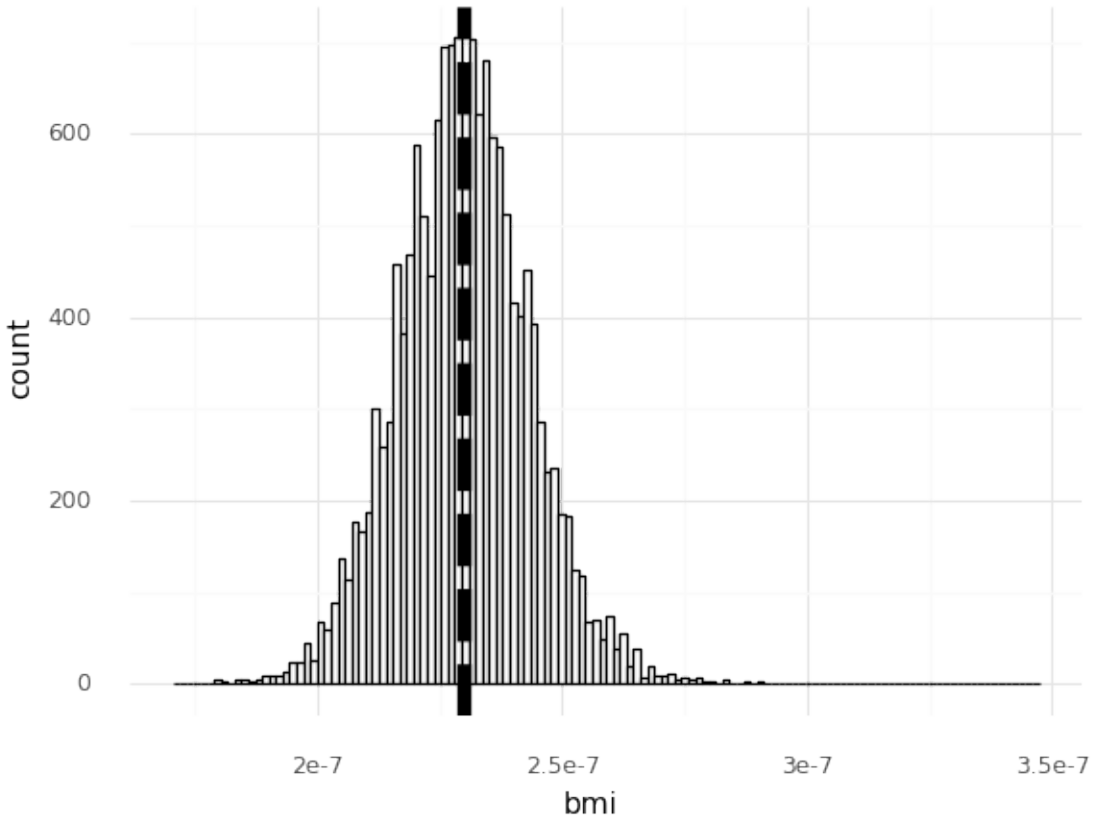
[27]: <ggplot: (8773494996363)>

[30]:
```
#7

ggplot(fifa, aes(x = "body_type", y = "weight_kg", fill = "body_type")) +␣
↪stat_summary(fun_data = "mean_sdl", geom = "bar")
```



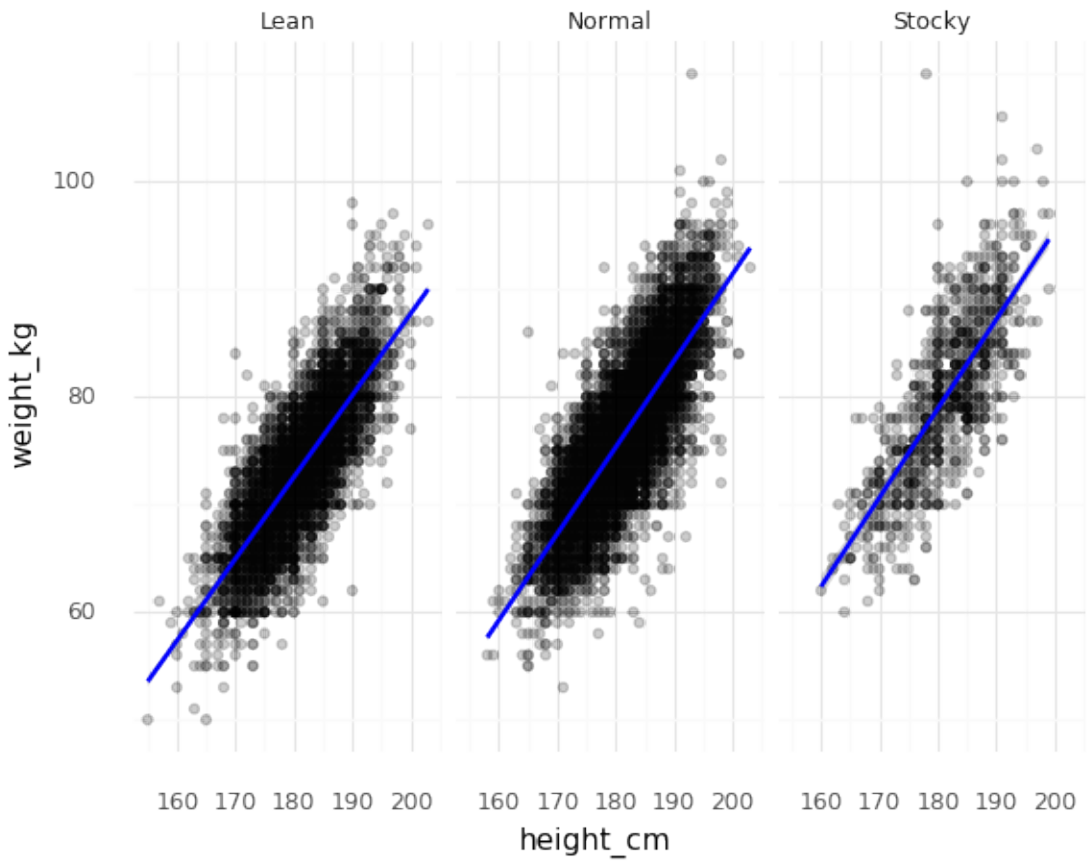
[30]: <ggplot: (8773494664183)>

[32]:
```
#8

height_sq = (fifa["height_cm"] * 100)**2

fifa["bmi"] = fifa["weight_kg"]/height_sq

(ggplot(fifa, aes(x = "bmi")) + geom_histogram(fill = "white", color = "black")␣
↪+
geom_vline(xintercept = fifa["bmi"].mean() , linetype = "dashed", size = 3) +
theme_minimal())
```



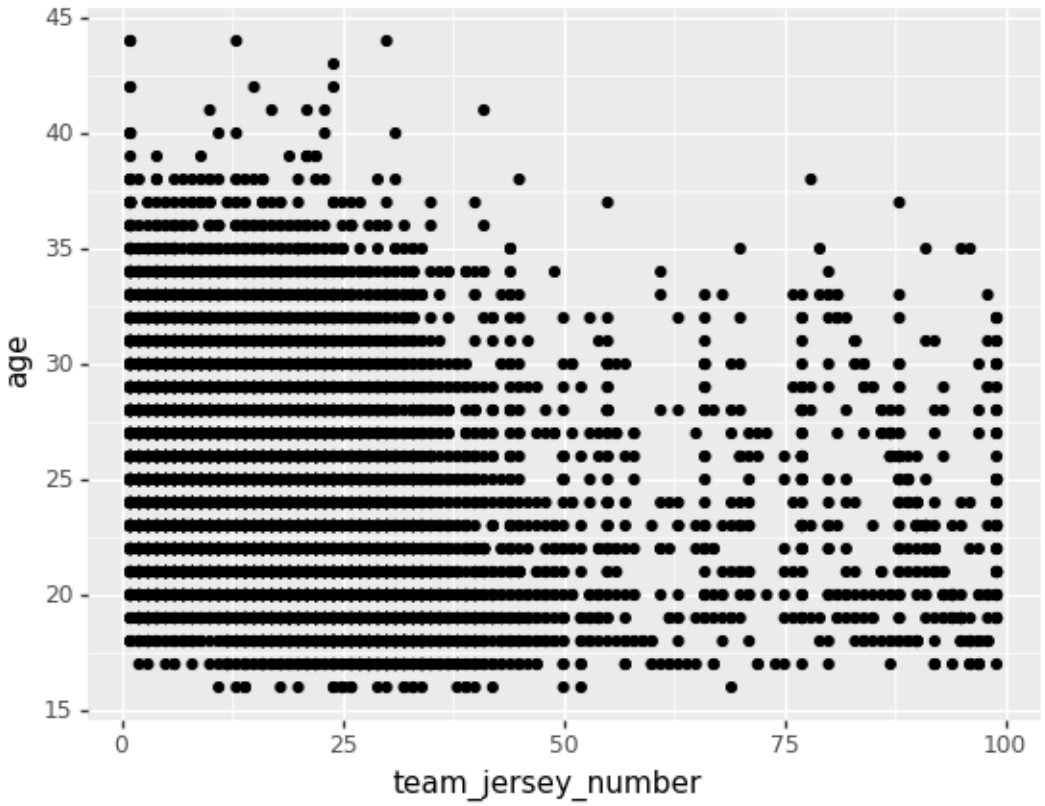
[32]: <ggplot: (8773494162122)>

[37]:
```
#9
(ggplot(fifa, aes(x = "height_cm", y = "weight_kg")) + geom_point(alpha = 0.2) +
 geom_smooth(method = "lm", color = "blue") +  facet_wrap("~body_type") +␣
 ↪theme_minimal())
```



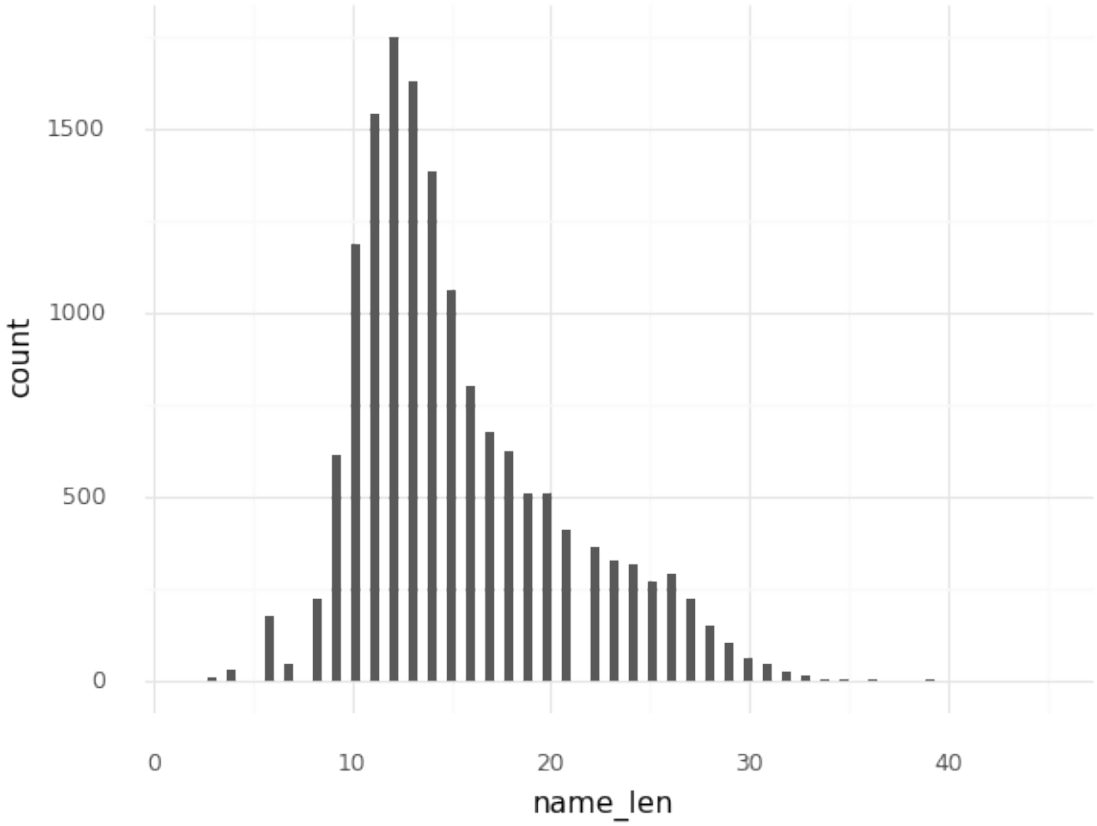
[37]: <ggplot: (8773540093118)>

[38]:
```
#10

ggplot(fifa, aes(x = "team_jersey_number", y = "age")) + geom_point()
```



[38]:
```
<ggplot: (8773542932041)>
```

[40]:
```
#11
# Let's use your CPSC230 skills, create a column in fifa called name_len that
# counts the number of characters in each player's long_name (spaces shouldn't␣
 ↪count).
# Then use ggplot/plotnine to create a histogram of name_len and add␣
 ↪+theme_minimal()
# to your graph. What is the typical range of name length?

fifa["name_len"] = [len("".join(name.split())) for name in fifa["long_name"]]

ggplot(fifa, aes(x = "name_len")) + geom_histogram() + theme_minimal()
```



[40]: <ggplot: (8773510756464)>

The typical range of name_len is between 5 and about 25 characters