Introduction to correlations

A/B TESTING IN R

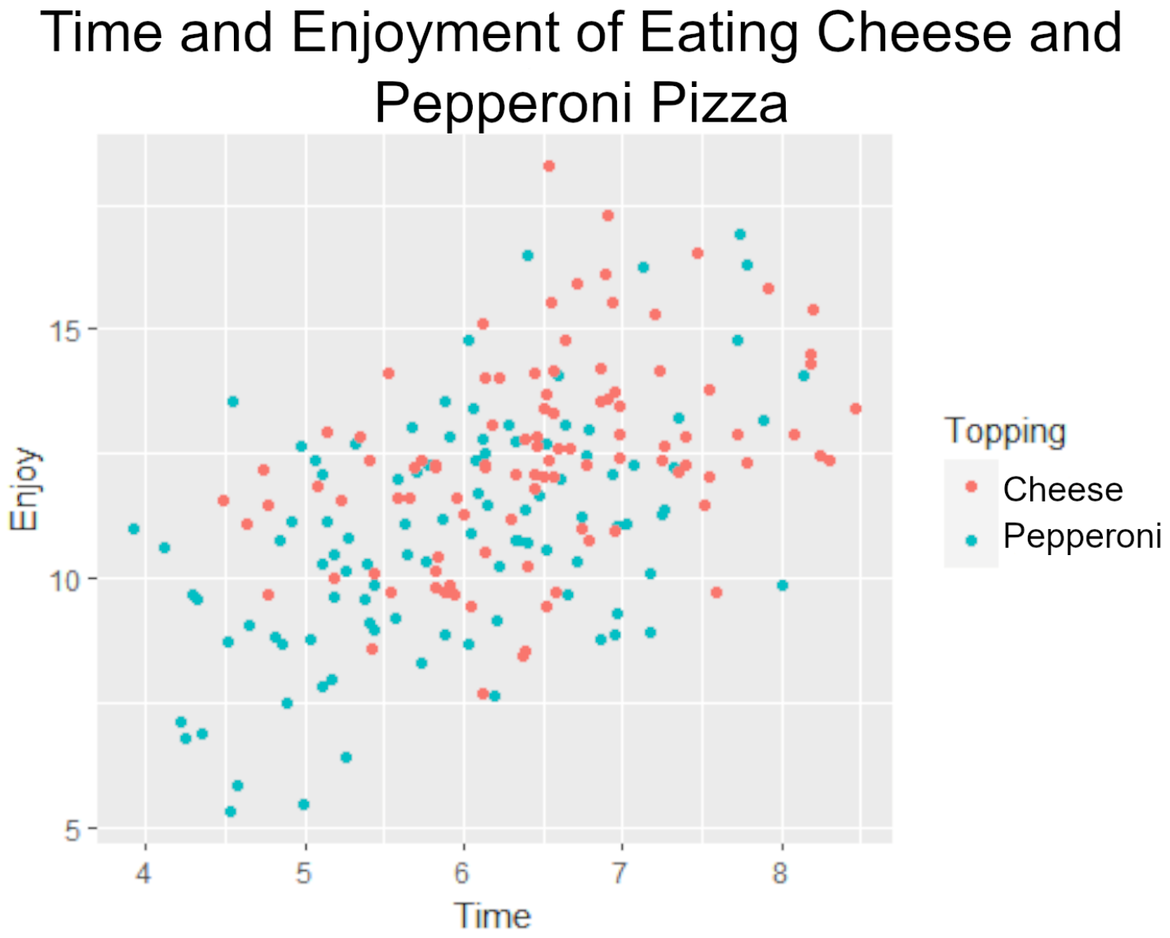


Lauryn Burleigh
Data Scientist

datacamp

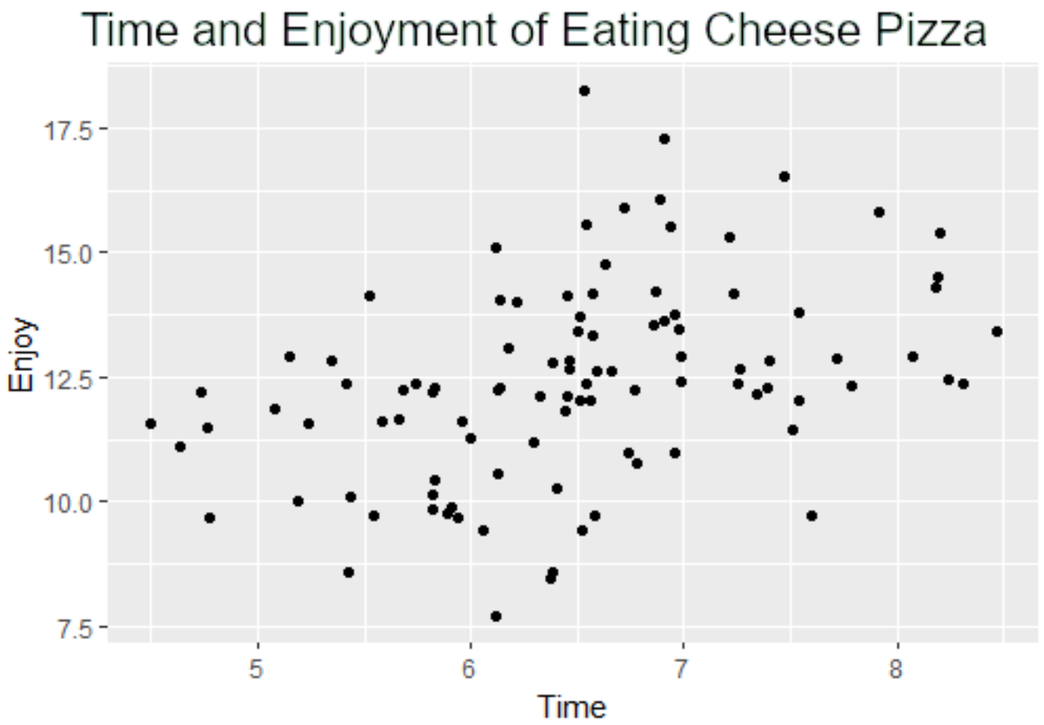
Correlation in A/B design

- Relationship strength and direction
- Two variables
- Increase or decrease in one variable per unit of the other variable
  - Enjoyment correlated with time to eat
- Ignore groups



Correlation in A/B design

- Within groups

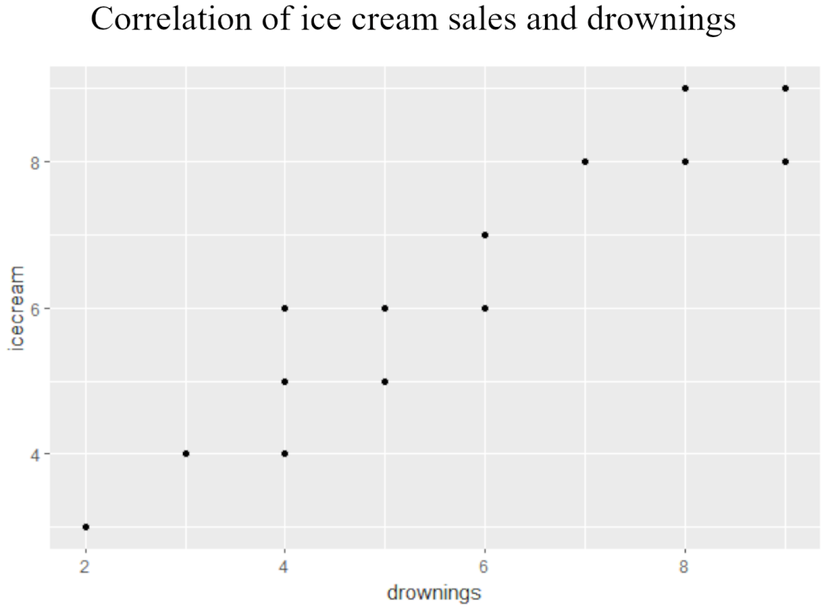




Correlation

- **Does NOT imply causation**
- Increase in drownings & ice cream sales
  - NOT causal
  - Likely: warm months
- Using A/B to deduce causation
  - Make and test changes
- A/B: cheese and pepperoni
  - Enjoyment and time to eat pizza
  - Group change - Relationship change
- Does not indicate dependency

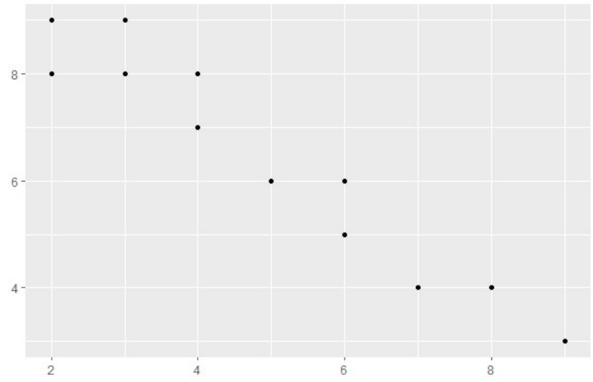
```
ggplot(data, aes(x = drownings,
                 y = icecream)) +
  geom_point()
```



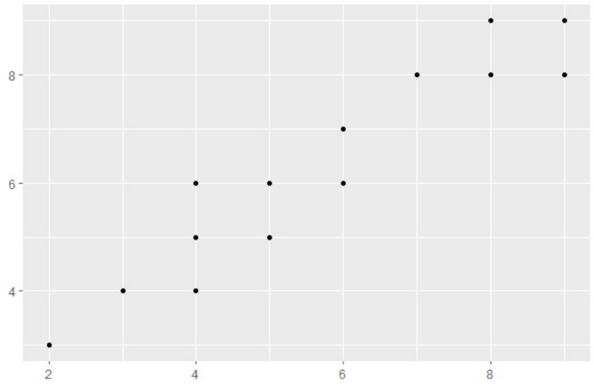
Correlation coefficient

- Correlation coefficient (r) - degree of association
  - -1 to +1
- More extreme values - stronger association
  - Better prediction

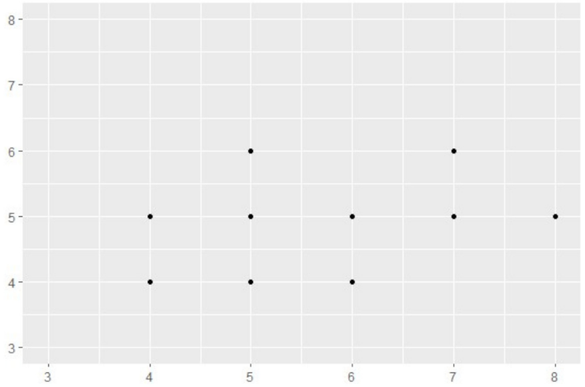
Negative correlation



Positive correlation



No significant correlation



Correlation values

- Proportion of variation in time eating that can be attributed to enjoyment
  - R^2

```
corvalue <- cor(data$time, data$enjoyment)
corvalue
```

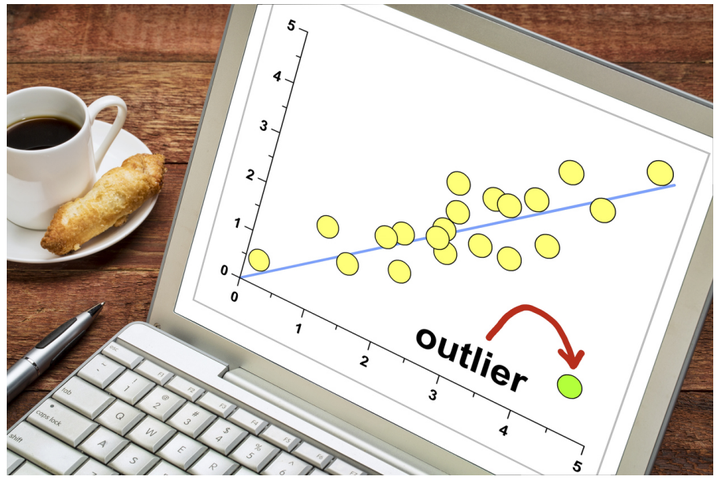
```
[1]    .73
```

```
corvalue^2
```

```
[1]        .5329
```

Correlation limitations

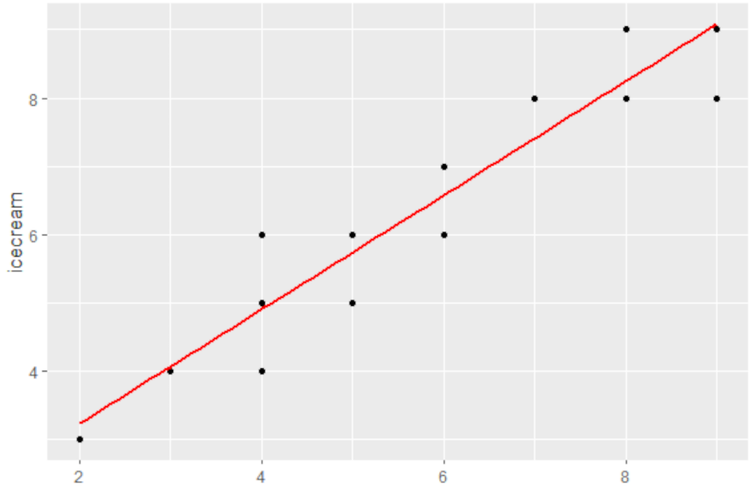
- Particularly susceptible to outliers



Correlation coefficients

- Not line of best fit
  - Regression
- Not indication of significance
  - Correlation coefficient and sample size for p-value

Positive correlation of ice cream sales and drownings



Let's practice!

A/B TESTING IN R

Pearson correlation

A/B TESTING IN R



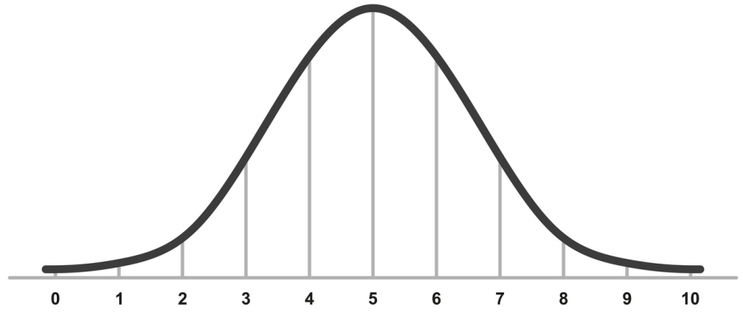
Lauryn Burleigh
Data Scientist

datacamp

Pearson correlation assumptions

- Choosing correlation test: properties of data
- Pearson correlation
  - Linear
  - Normal distribution





Pearson correlation and A/B tests

- Group A: Cheese pizza
- Group B: Pepperoni pizza

- Relationship of time to eat pizza and enjoyment

Ignoring groups

- Null hypothesis - no relationship between time to eat and enjoyment of the pizza

Each group

- Null hypothesis - no relationship between time to eat and enjoyment of the Cheese pizza
- Null hypothesis - no relationship between time to eat and enjoyment of the Pepperoni pizza

Determine sample size

- `r` expected effect (found with `cor()`)
- `sig.level` at which we reject the null

```
library(pwr)
pwr.r.test(r = 0.3, power = 0.80,
           sig.level = 0.05)
```

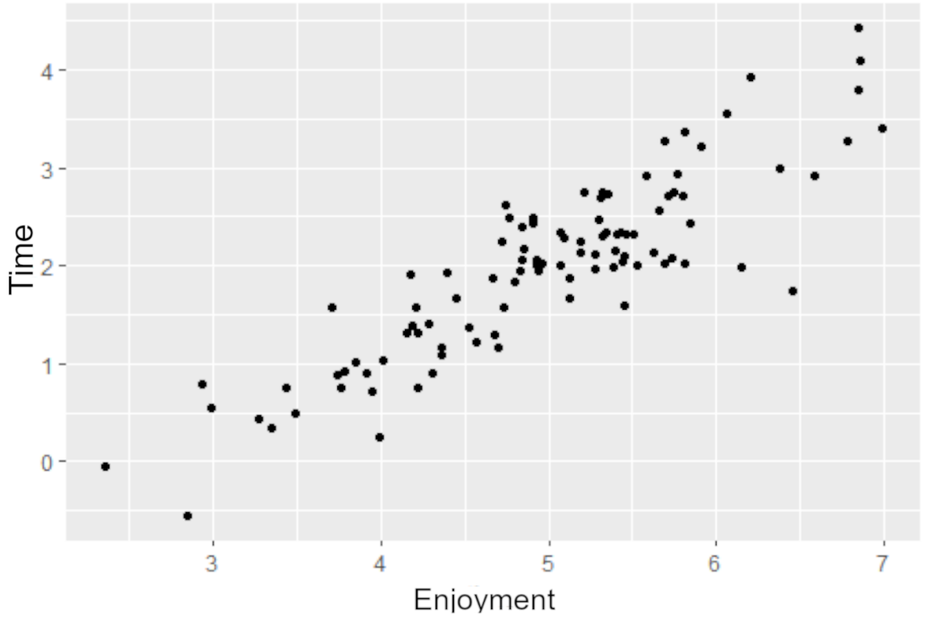
```
     approximate correlation power
 calculation (arctangh transformation)

              n = 84.07364
              r = 0.3
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

Assessing linearity

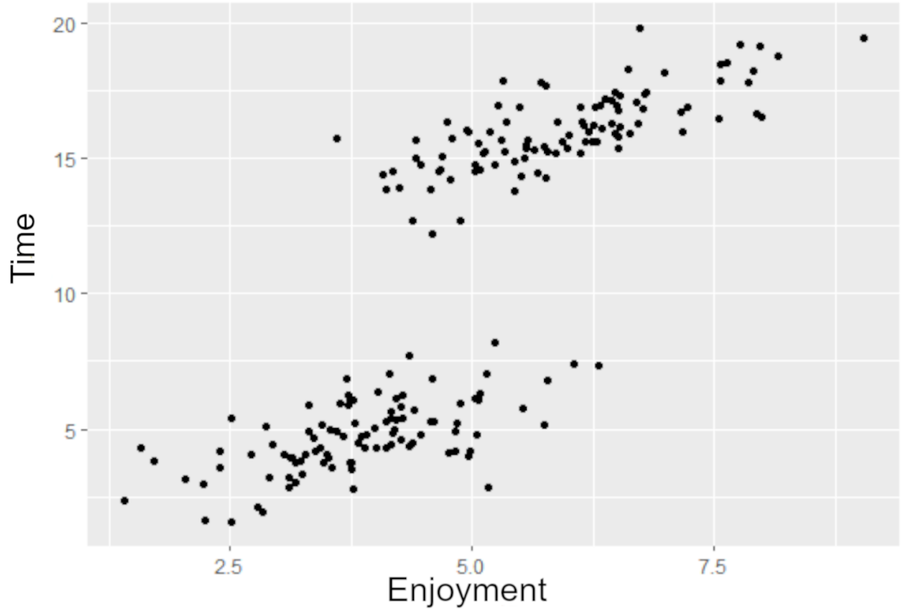
```
ggplot(pizza, aes(x = enjoyment,
                  y = time)) +
    geom_point()
```

Correlation of pizza enjoyment and time to eat




```
ggplot(pizza, aes(x = enjoyment,
                  y = time)) +
    geom_point()
```

Distinctive correlation clusters of pizza enjoyment and time to eat



Assessing normality

```
shapiro.test(pizza$time)
```

```
    Shapiro-Wilk normality test
data:  pizza$time
W = 0.98686, p-value = 0.4282
```

- Data is normal

```
shapiro.test(pizza$enjoyment)
```

```
    Shapiro-Wilk normality test
data:  pizza$enjoyment
W = 0.98916, p-value = 0.5971
```

- Data is normal

Pearson ignoring groups

```
cor.test(~ time + enjoyment,
         data = pizza,
         method = "pearson")
```

```
Pearson's product-moment correlation

data:  time and enjoyment
t = 22.304, df = 88, p-value = 0.0218
alternative hypothesis: true correlation
is not equal to 0
95 percent confidence interval:
 0.8833166 0.9479256
sample estimates:
      cor
0.3021878
```

Proportion of variance: cor^2

```
0.30^2
```

```
[1] 0.09
```

Pearson within groups

```
cor.test(~ time + enjoyment,
         data = pizza,
         subset =
             (Topping == "Cheese"),
         method = "pearson")
```

```
Pearson's product-moment correlation

data:  Time and Enjoy
t = 11.121, df = 98, p-value < 2.2e-16
alternative hypothesis: true correlation
is not equal to 0
95 percent confidence interval:
 0.6451710 0.8226595
sample estimates:
      cor
0.746935
```

Power analysis

Pearson ignoring groups:

```
Pearson's product-moment correlation

data:  time and enjoyment
t = 22.304, df = 88, p-value = 0.0218
alternative hypothesis: true
correlation is not equal to 0
95 percent confidence interval:
 0.8833166 0.9479256
sample estimates:
      cor
0.3021878
```

```
library(pwr)
pwr.r.test(r = 0.302, n = 100,
           sig.level = 0.022)
```

```
     approximate correlation power
  calculation (arctangh transformation)

              n = 100
              r = 0.302
      sig.level = 0.022
          power = 0.7853514
    alternative = two.sided
```

Let's practice!

A/B TESTING IN R

Spearman-rank correlation

A/B TESTING IN R



Lauryn Burleigh
Data Scientist

datacamp

Spearman correlation assumptions

- Rank ordered - ordered by position
- High Spearman correlation - similar positions in the two variables
- Continuous or discrete data
  - Ordinal, interval, or ratio

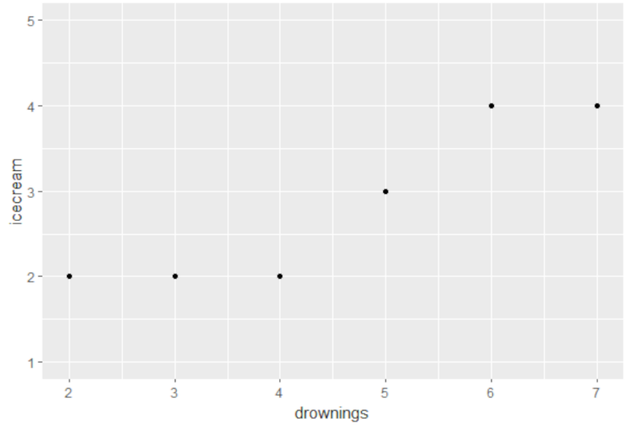


Monotonic relationships

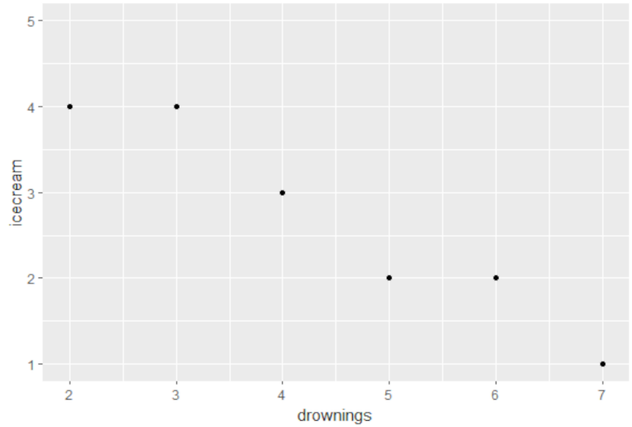
- Spearman correlation (rho) - monotonic relationship
  - Not consistently increasing or decreasing
- Monotonic less restrictive than linear

```
ggplot(data, aes(x = drownings,
                 y = icecream)) +
  geom_point()
```

Positive monotonic relationship between ice cream and drownings



Negative monotonic relationship between ice cream and drownings



Hypothesis and sample size

- Null hypothesis - no monotonic association between time and enjoyment of pizza

```
library(pwr)
pwr.r.test(r = 0.3, power = 0.80,
           sig.level = 0.05)
```

```
     approximate correlation power
 calculation (arctangh transformation)

              n = 84.07364
              r = 0.3
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

Spearman ignoring groups

```
cor.test(~ enjoyment + time,
         data = pizza,
         method = "spearman",
        exact = FALSE)
```

```
    Spearman's rank correlation rho
data:  time and enjoyment
S = 2, p-value = .003245
alternative hypothesis: true rho is
not equal to 0
sample estimates:
      rho
0.9984962
```

```
samp <- length(pizza$time)
```

```
[1]  90
```

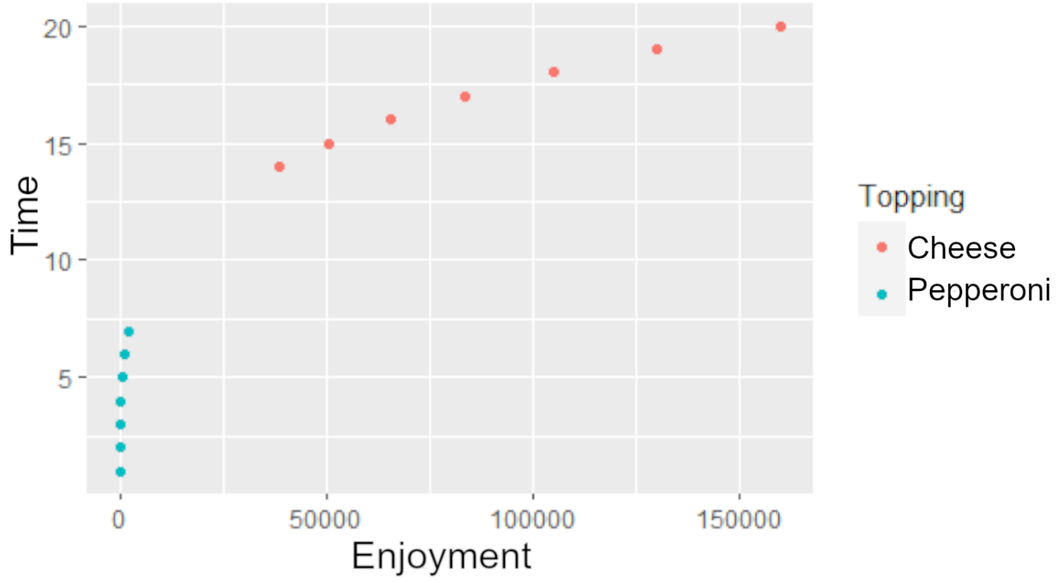
Spearman within groups

```
cor.test(~ enjoyment + time,
         data = pizza,
         subset =
             (Topping == "Cheese"),
         method = "spearman",
         exact = FALSE)
```

```
    Spearman's rank correlation rho
data:  time and enjoyment
S = 1.2434e-14, p-value = 0.0003968
alternative hypothesis: true rho is
not equal to 0
sample estimates:
rho
  1
```

```
ggplot(pizza, aes(x = enjoyment,
                  y = time,
                  color = Topping)) +
  geom_point()
```

Monotonic relationship across groups
Linear relationship within groups



Spearman power analysis

```
library(pwr)
pwr.r.test(r = 0.998, n = 90,
           sig.level = 0.003)
```

```
     approximate correlation power
calculation (arctangh transformation)

              n = 90
              r = 0.998
      sig.level = 0.003
          power = 1
    alternative = two.sided
```

Referring to the output

```
rhotest <- cor.test(~ enjoyment + time,
         data = pizza,
         method = "spearman")
samp <- length(pizza$time)
```

```
library(pwr)
pwr.r.test(r = rhotest$estimate,
           sig.level = rhotest$p.value,
           n = samp)
```

```
     approximate correlation power
calculation (arctangh transformation)

              n = 90
              r = 0.998
      sig.level = 0.003
          power = 1
    alternative = two.sided
```

Let's practice!

A/B TESTING IN R

Reporting A/B test results

A/B TESTING IN R



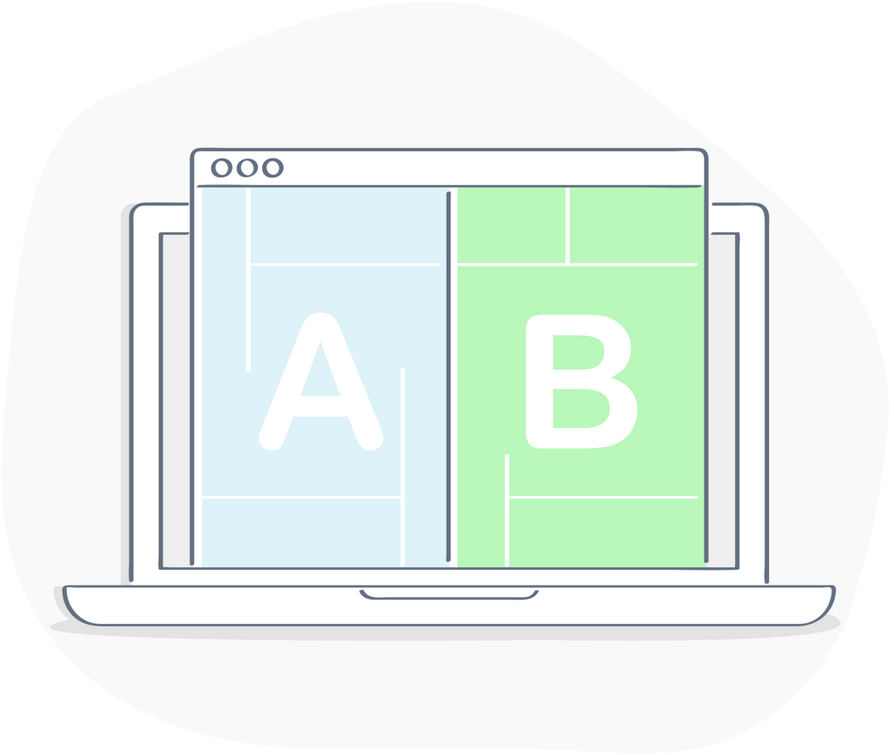
Lauryn Burleigh
Data Scientist

datacamp

A/B visualizations

- Send conclusions outside the data team

- Clear visualizations
- Accurately convey findings



Bar plots

- Hypothesis comparing group means
  - t-test

Visualizes:

- Mean
- Error bars
  - Standard deviation
  - Standard error
  - Confidence interval



Bar data formatting

```
library(dplyr)

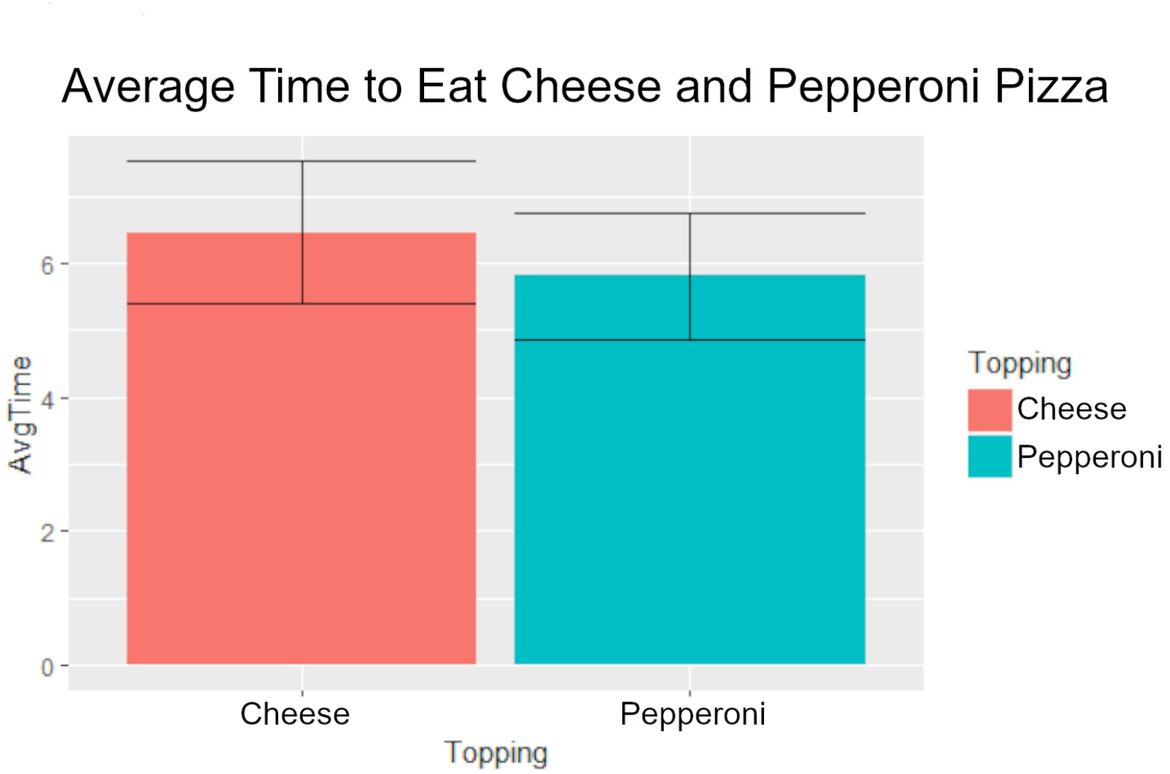
data <- pizza %>%
    group_by(Topping) %>%
    summarise(AvgTime = mean(Time),
              SdTime = sd(Time))
```

```
data
```

```
# A tibble: 2 × 3
  Topping   AvgTime SdTime
  <chr>       <dbl>  <dbl>
1 Cheese       5.81  0.948
2 Pepperoni    6.46  1.06
```

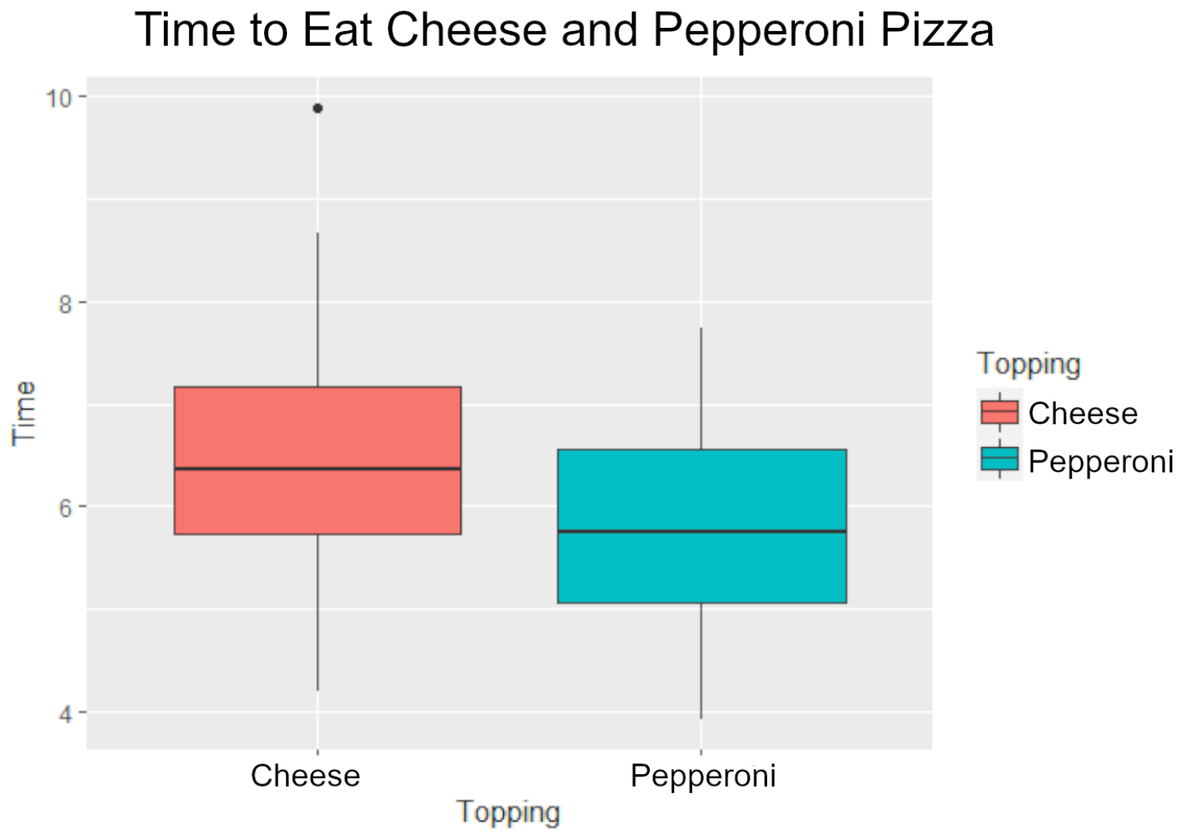
Using bar plots

```
library(ggplot2)
ggplot(data, aes(x = Topping,
                 y = AvgTime,
                 fill = Topping)) +
  geom_bar(stat =  "summary",
           fun = "mean") +
  geom_errorbar(aes(
    ymin = AvgTime - SdTime,
    ymax = AvgTime + SdTime))
```



Box plots

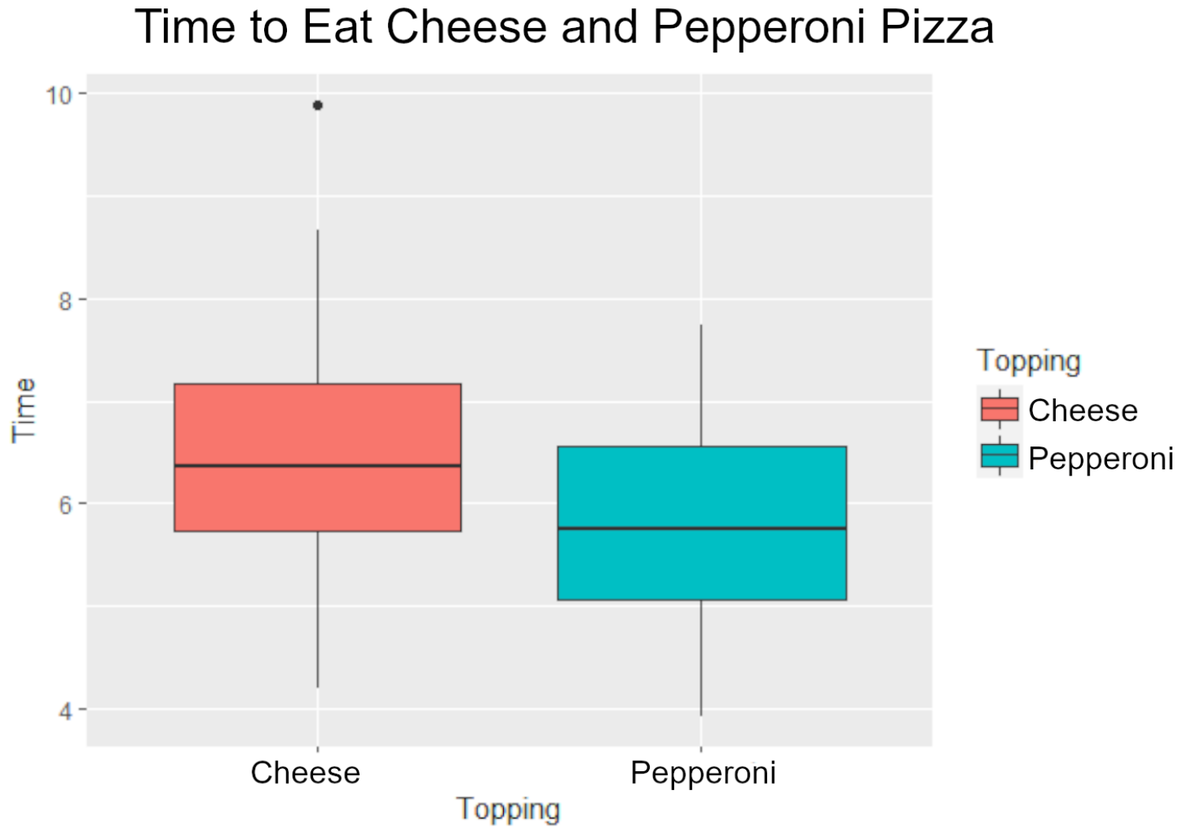
- Comparing group medians
  - Mann-Whitney U
  - Chi-square
  - Fisher's Exact
  - t-tests
- Bottom box limit: 25% of data points below
- Top box limit: 75% of data points below
- Whiskers: highest and lowest values
  - Outliers excluded
- Outliers: point outside of whiskers



Using box plots

- Data in long format
- Ideal for researchers or statisticians or assessing non-normal distributions

```
ggplot(pizza, aes(x = Topping, y = Time,
                  fill = Topping)) +
  geom_boxplot()
```



Scatter plots

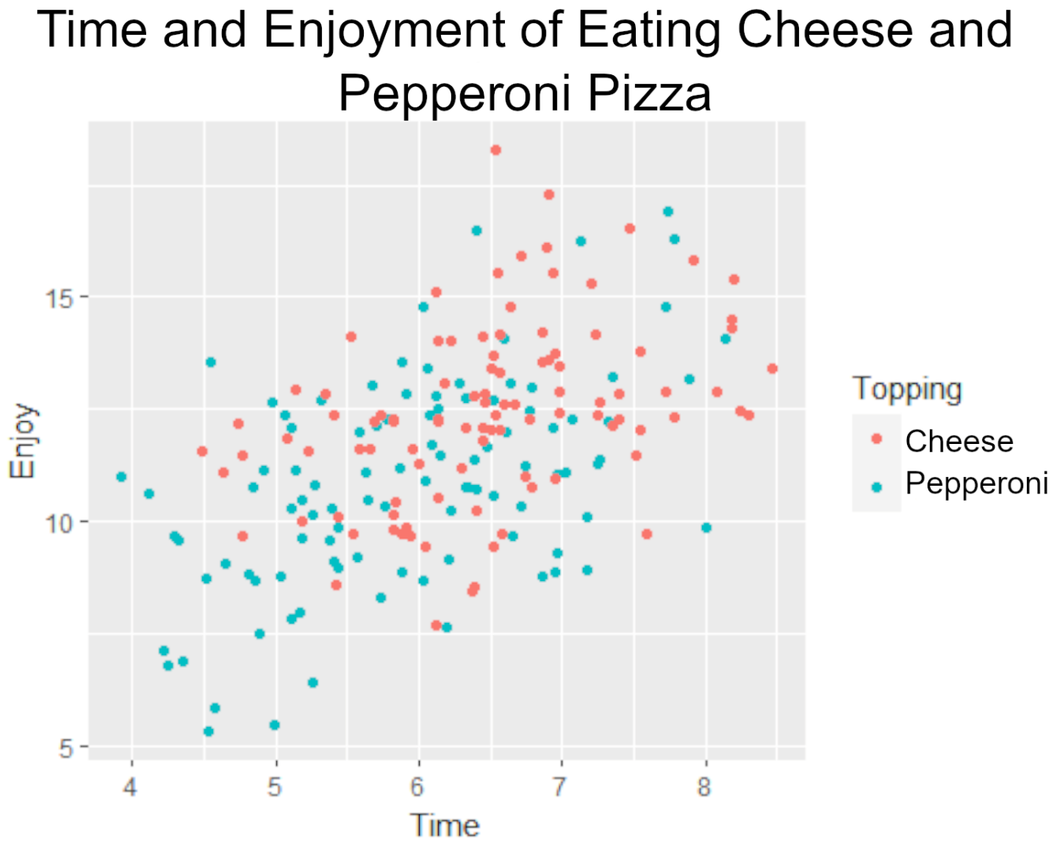
- Relationship between variables
  - Correlations
- Every point
- Trend: direction and rate



Using scatter plots

```
library(ggplot2)

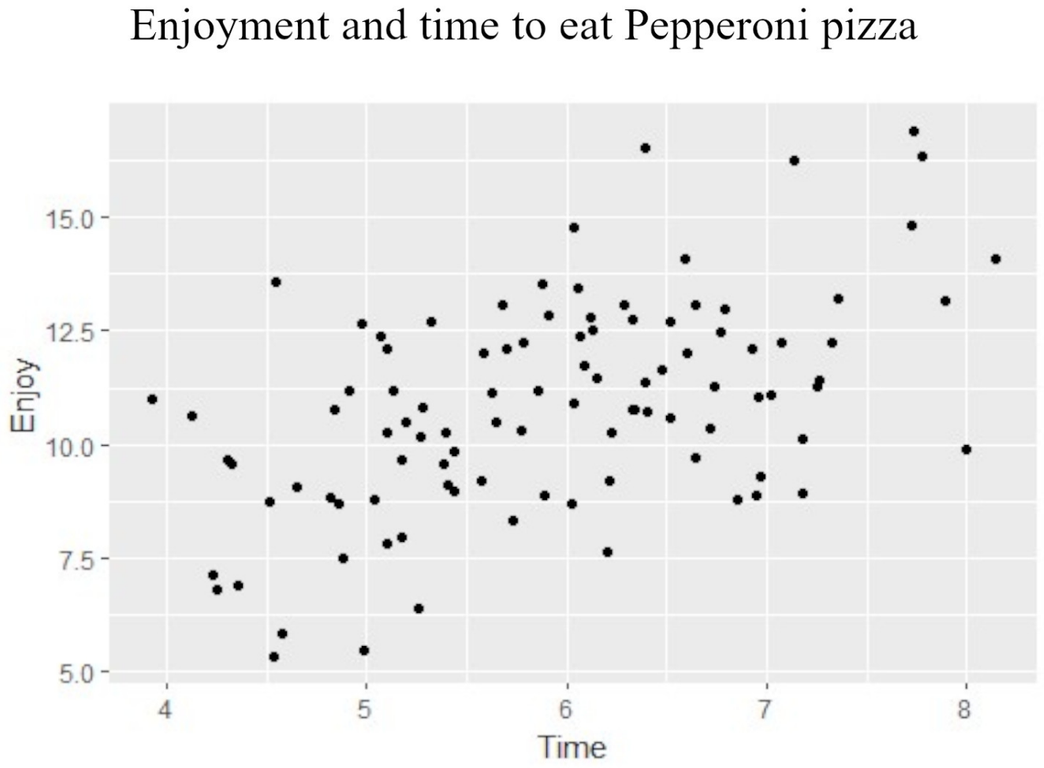
ggplot(pizza, aes(x = Time,
                  y = Enjoyment,
                  fill = Topping)) +
  geom_point()
```



Using scatter plots

- Subset data to assess one group

```
ggplot(subset(pizza,
              Topping %in%
              "Pepperoni"),
       aes(x = Time, y = Enjoy)) +
  geom_point()
```



Presenting data

Audience has no statistics background:

- Statistic values uninformative
- Use real-world findings
- Test conclusion
- Power analysis
- Error bars



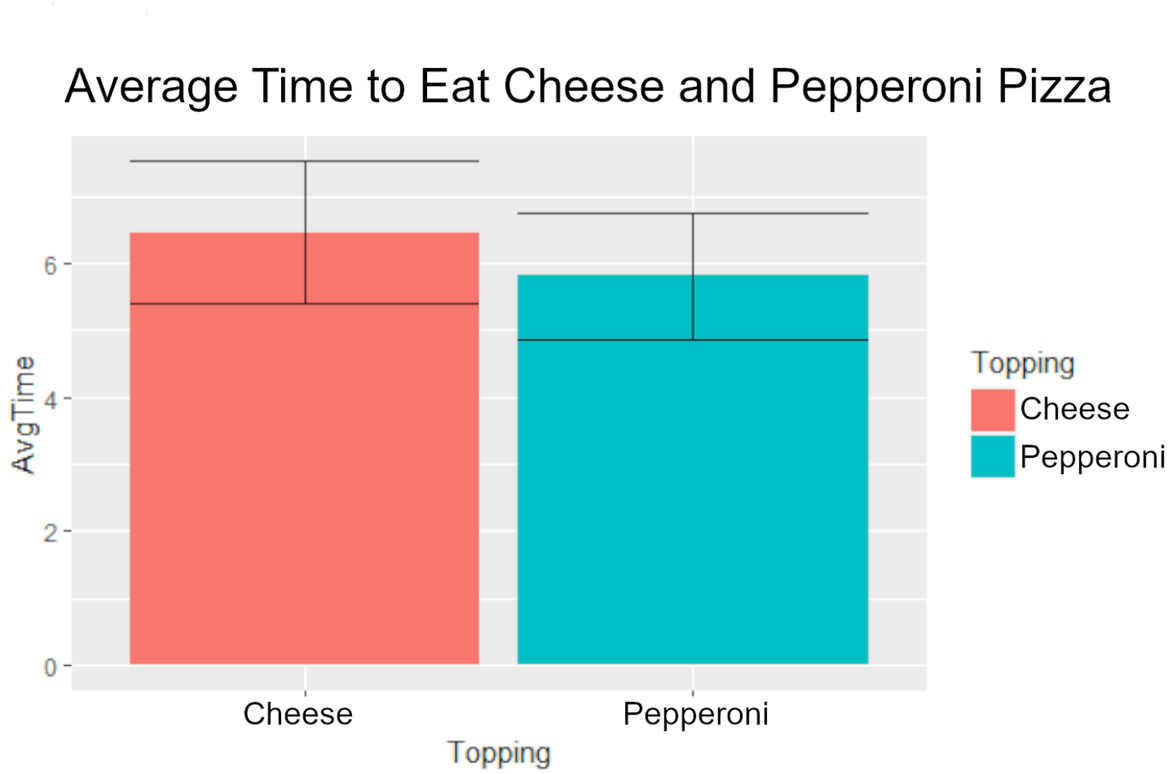
Presenting data example

Results:

- Significant t-test
- Power = 0.9

Presentation:

- Significant difference was found
- Pepperoni pizza was eaten quicker than Cheese
- Reliable results
- No large variation (error bars)



Let's practice!

A/B TESTING IN R