Statistics and Data Analysis Assignment

Domen Mohorčič, Larsen Cundrič

09/04/2021

Roller Coaster Dataset Analysis

State hypothesis here...

Basic analysis and plots

```
roller_coasters_raw <- readr::read_csv('datasets/roller_coasters.csv')
```

```
## Parsed with column specification:
## cols(
##   Name = col_character(),
##   Park = col_character(),
##   City = col_character(),
##   State = col_character(),
##   Country = col_character(),
##   Type = col_character(),
##   Construction = col_character(),
##   Height = col_double(),
##   Speed = col_double(),
##   Length = col_double(),
##   Inversions = col_character(),
##   Numinversions = col_double(),
##   Duration = col_double(),
##   GForce = col_double(),
##   Opened = col_double(),
##   Region = col_character()
## )
```

```
# GForce to many missing values..
#
summary(roller_coasters_raw)
```

```
##      Name               Park               City              State
##  Length:408         Length:408         Length:408         Length:408
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
```

```
## 
## 
##     Country              Type             Construction           Height
## Length:408          Length:408          Length:408          Min.   :  2.438
## Class :character    Class :character    Class :character    1st Qu.:  8.651
## Mode  :character    Mode  :character    Mode  :character    Median : 18.288
##                                                             Mean   : 23.125
##                                                             3rd Qu.: 33.167
##                                                             Max.   :128.016
##                                                             NA's   :82
##      Speed              Length           Inversions        Numinversions
## Min.   :  9.72     Min.   :  12.19     Length:408          Min.   : 0.0000
## 1st Qu.: 45.00     1st Qu.: 291.00     Class :character    1st Qu.: 0.0000
## Median : 68.85     Median : 415.75     Mode  :character    Median : 0.0000
## Mean   : 69.36     Mean   : 597.04                         Mean   : 0.7843
## 3rd Qu.: 88.95     3rd Qu.: 833.12                         3rd Qu.: 0.0000
## Max.   :194.40     Max.   :2243.02                         Max.   :10.0000
## NA's   :138        NA's   :90
##     Duration          GForce             Opened            Region
## Min.   :  0.3     Min.   :2.100      Min.   :1924       Length:408
## 1st Qu.: 75.0     1st Qu.:3.175      1st Qu.:1991       Class :character
## Median :108.0     Median :4.500      Median :1999       Mode  :character
## Mean   :112.5     Mean   :4.115      Mean   :1995
## 3rd Qu.:140.8     3rd Qu.:5.000      3rd Qu.:2004
## Max.   :300.0     Max.   :6.200      Max.   :2014
## NA's   :216       NA's   :348        NA's   :28
```
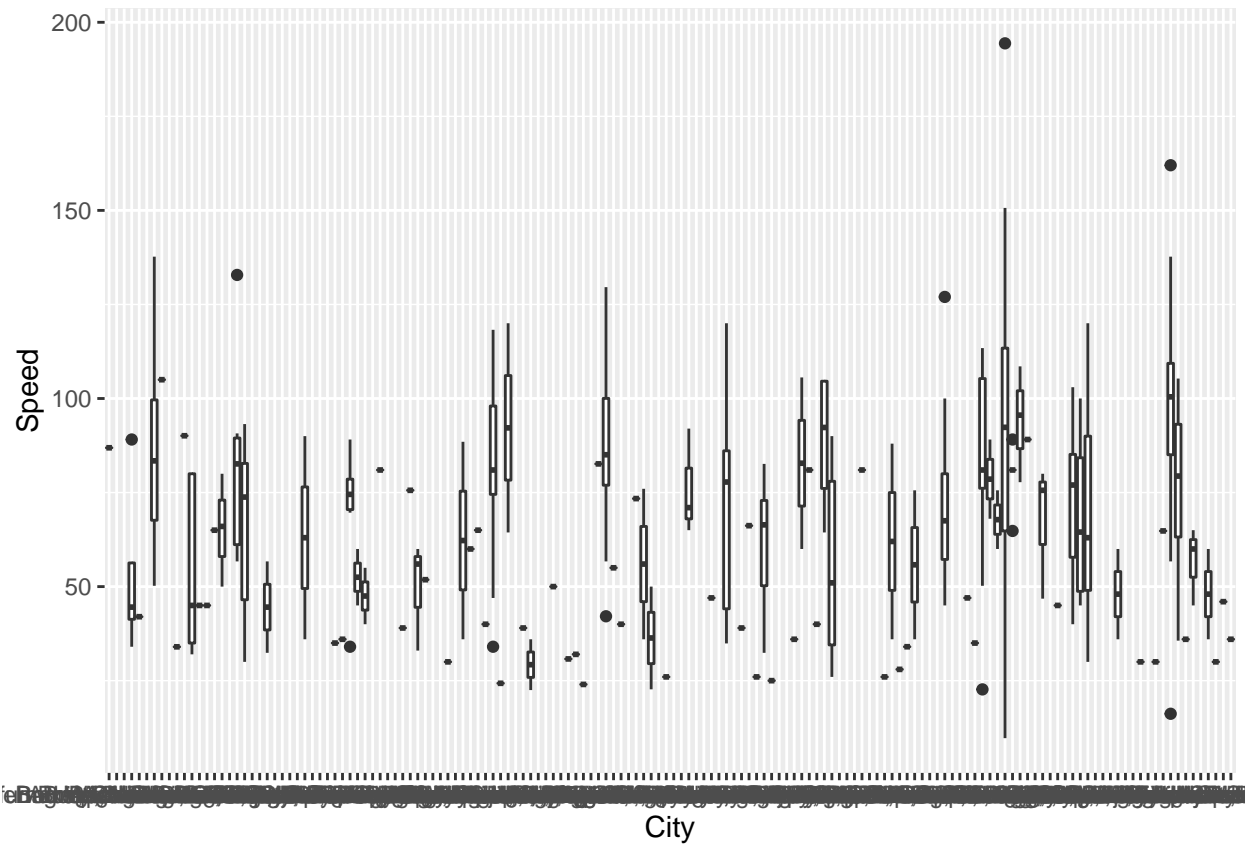
```r
ggplot(data = roller_coasters_raw) +
  geom_boxplot(mapping = aes(x = Country, y = Speed))
```

```
## Warning: Removed 138 rows containing non-finite values (stat_boxplot).
```
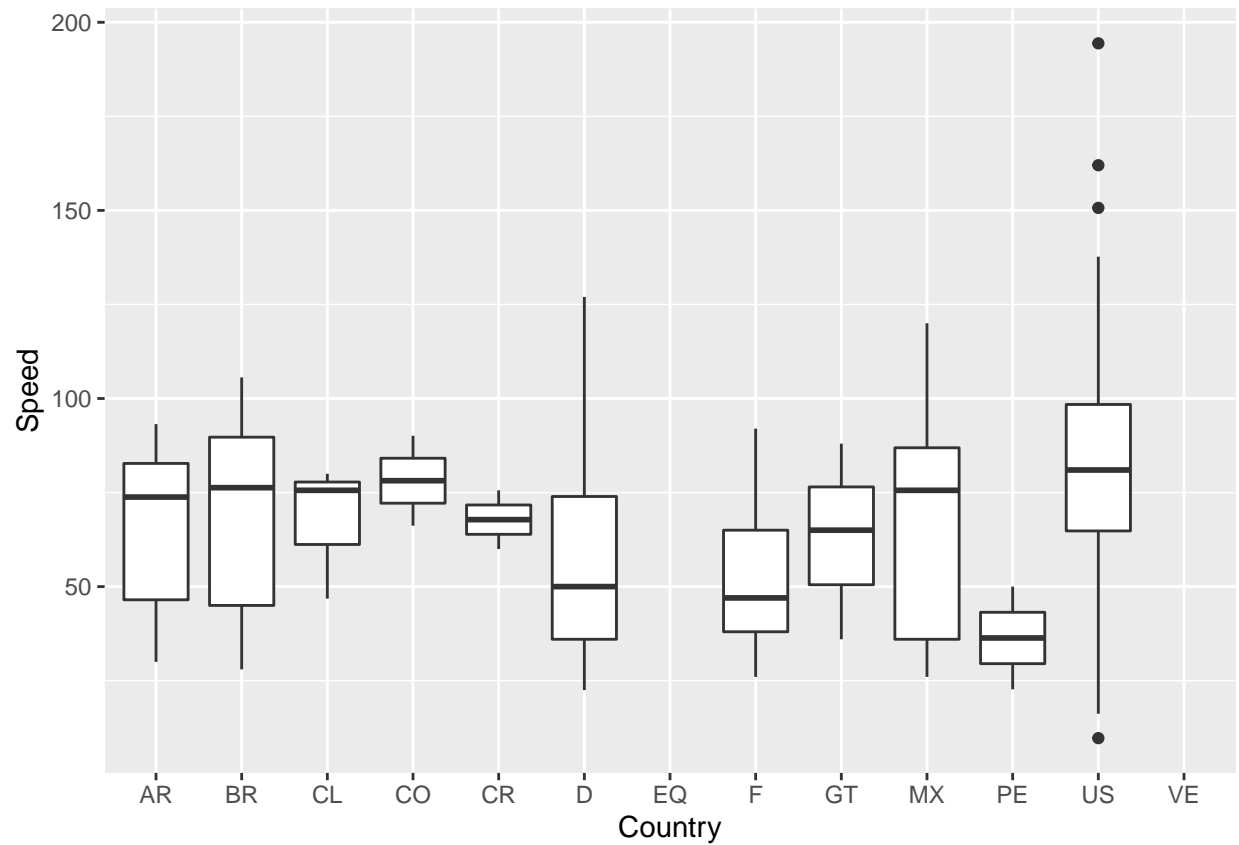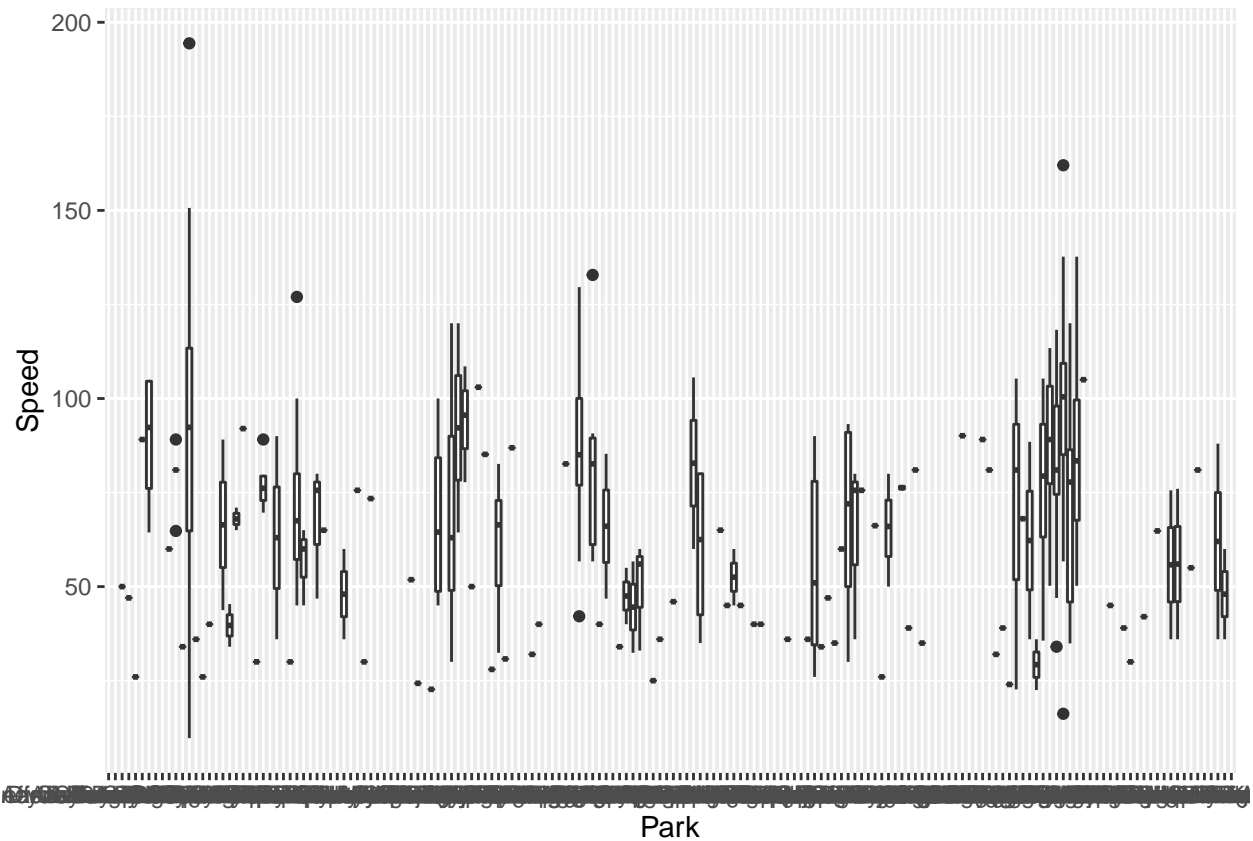
```
ggplot(data = roller_coasters_raw) +
  geom_boxplot(mapping = aes(x = City, y = Speed))
```

```
## Warning: Removed 138 rows containing non-finite values (stat_boxplot).
```
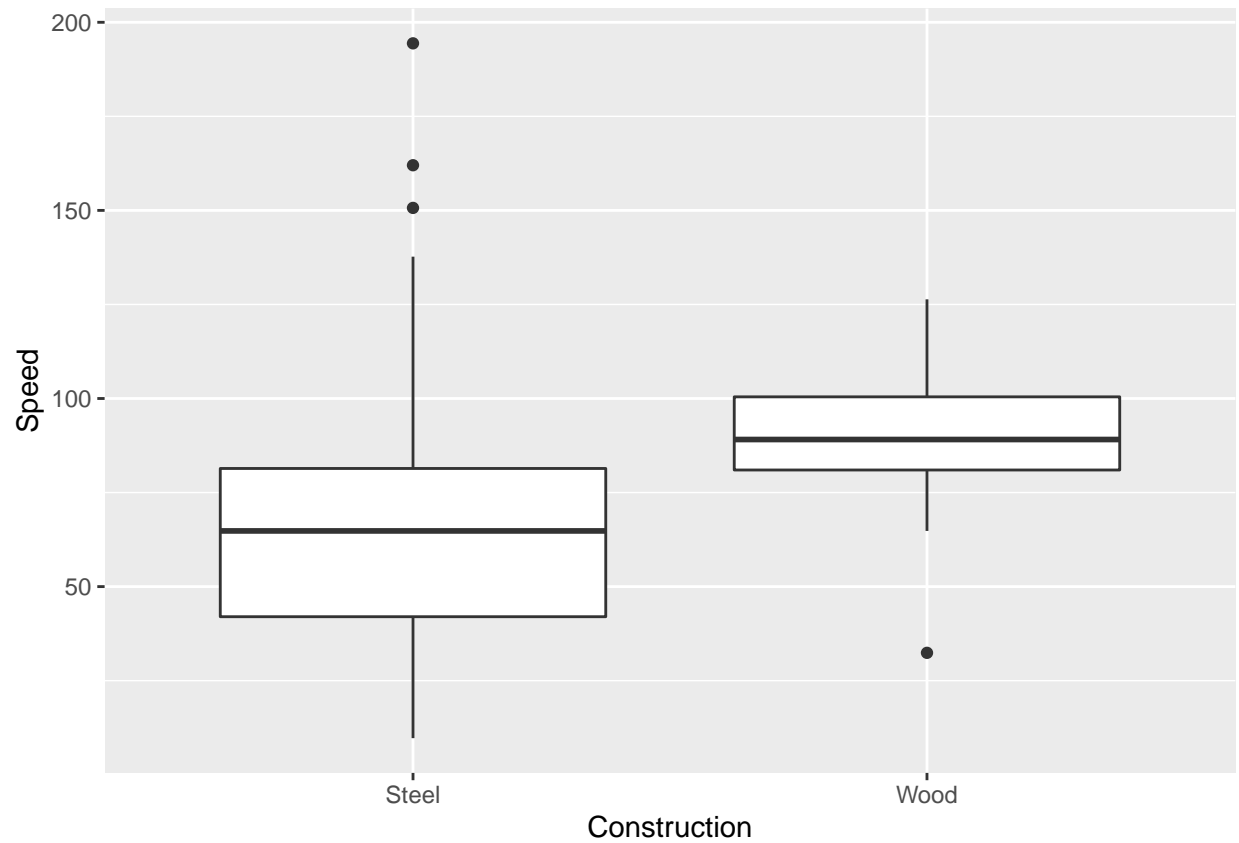
```
ggplot(data = roller_coasters_raw) +
  geom_boxplot(mapping = aes(x = Park, y = Speed))
```

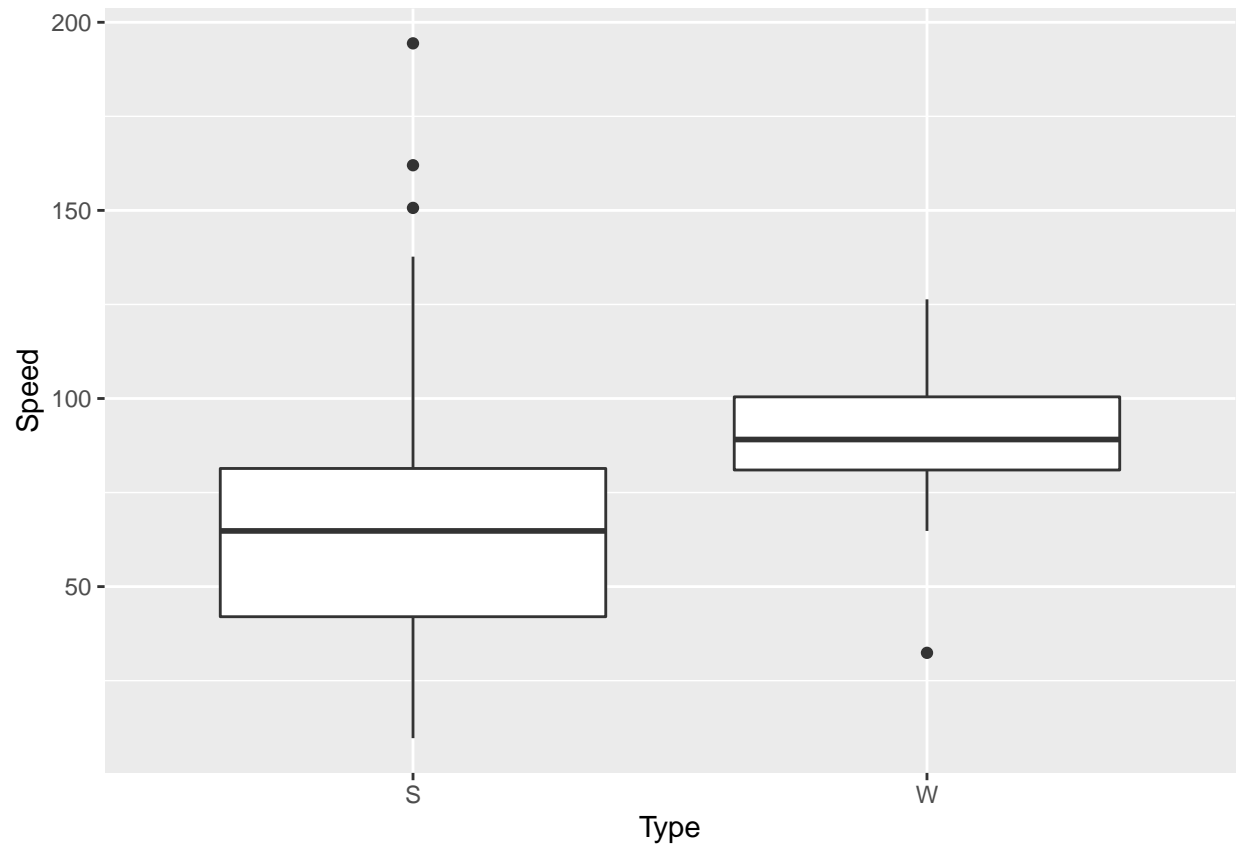## Warning: Removed 138 rows containing non-finite values (stat_boxplot).

```
ggplot(data = roller_coasters_raw) +
  geom_boxplot(mapping = aes(x = Construction, y = Speed))
```

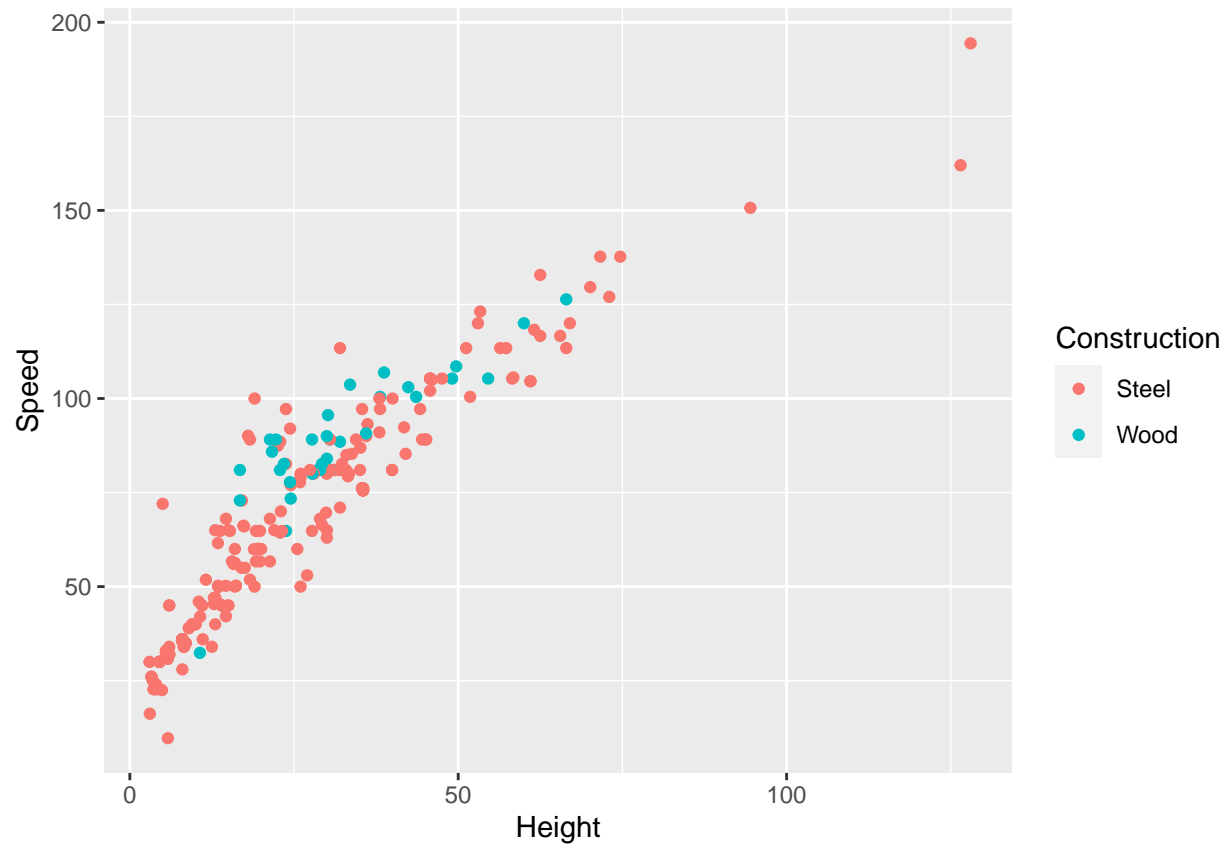## Warning: Removed 138 rows containing non-finite values (stat_boxplot).

```
# Type is same as Construction?
ggplot(data = roller_coasters_raw) +
  geom_boxplot(mapping = aes(x = Type, y = Speed))
```

Warning: Removed 138 rows containing non-finite values (stat_boxplot).
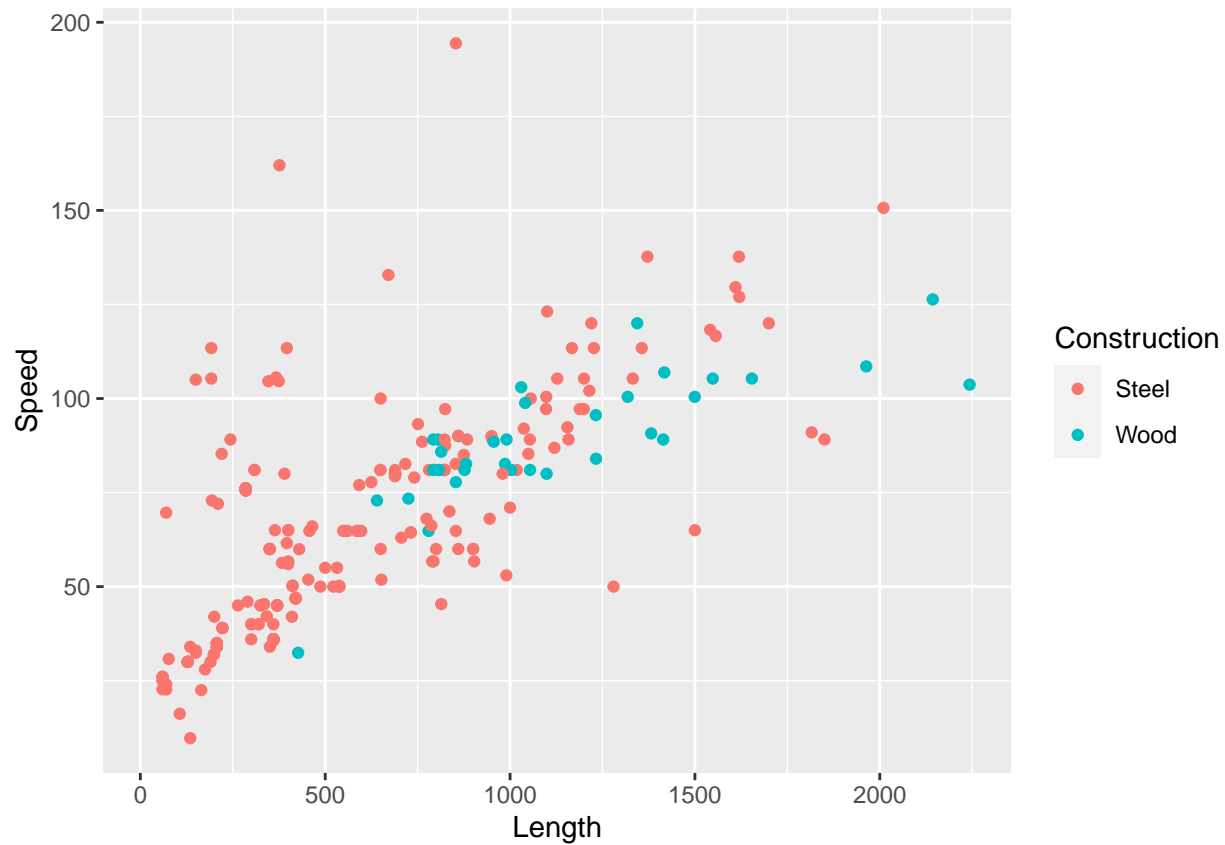
```
roller_coasters_raw %>%
  ggplot() +
    geom_point(aes(x = Height, y = Speed, color = Construction))
```

```
## Warning: Removed 150 rows containing missing values (geom_point).
```
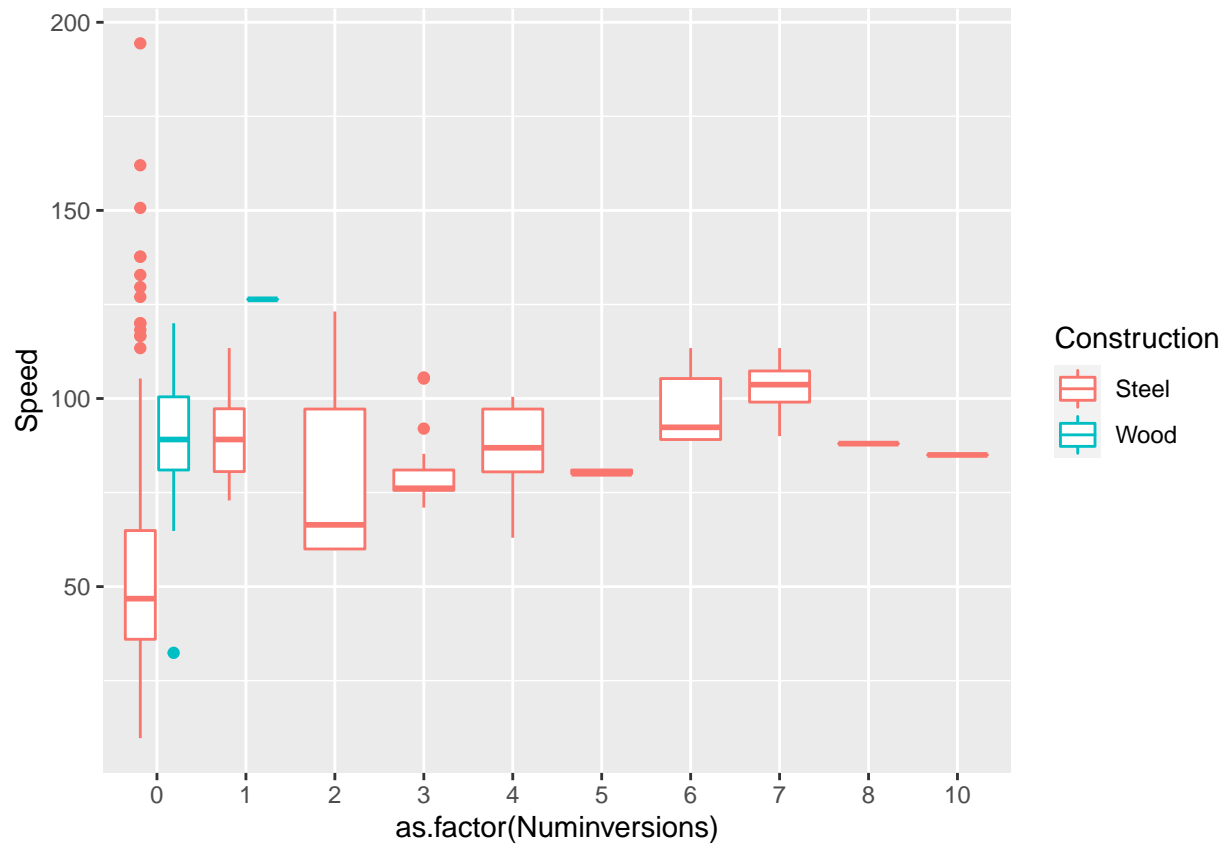
```
roller_coasters_raw %>%
  ggplot() +
    geom_point(aes(x = Length, y = Speed, color = Construction))
```

Warning: Removed 148 rows containing missing values (geom_point).

```
roller_coasters_raw %>%
  ggplot() +
    geom_boxplot(aes(x = as.factor(Numinversions), y = Speed, color = Construction))
```
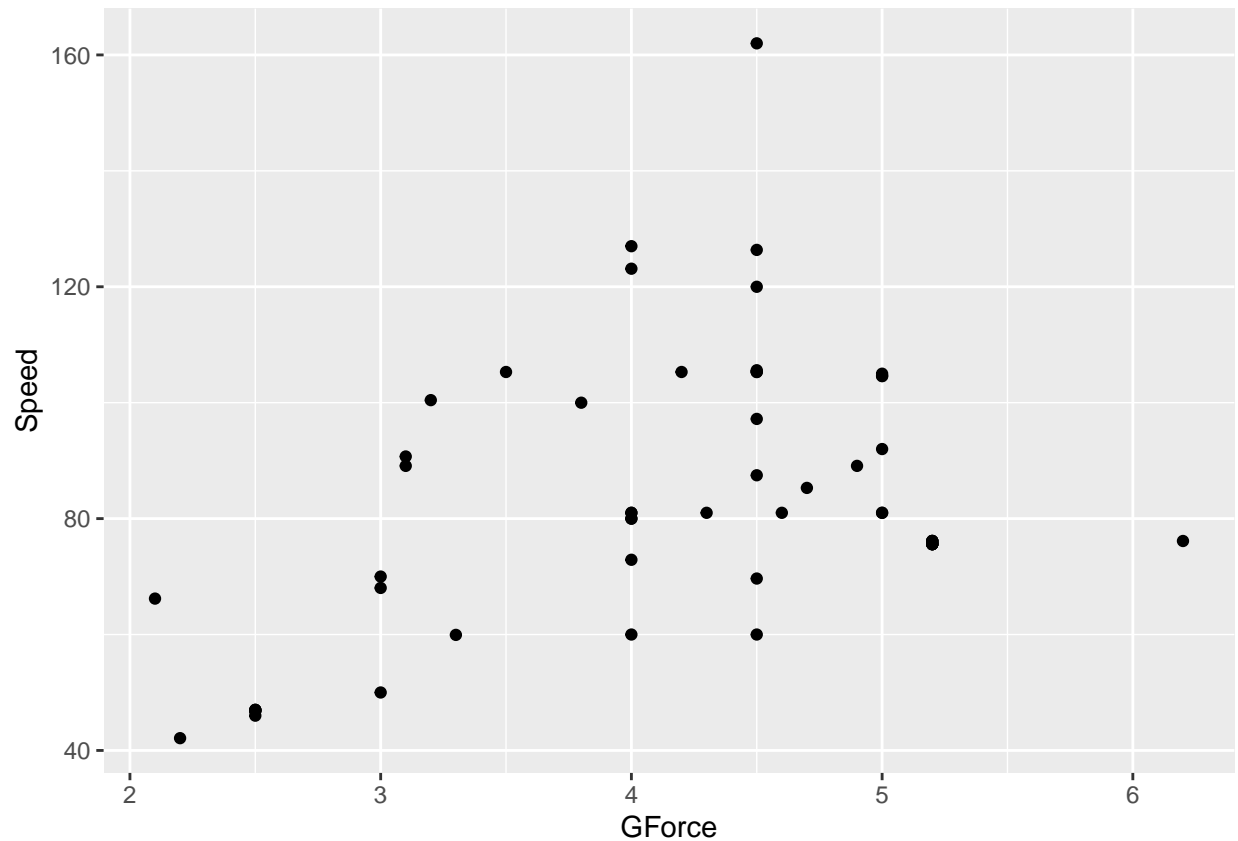
Warning: Removed 138 rows containing non-finite values (stat_boxplot).

```
# dost lame ... lahko spustimo
roller_coasters_raw %>%
  filter(!is.na(GForce)) %>%
  ggplot() +
    geom_point(aes(x = GForce, y = Speed))
```

Warning: Removed 2 rows containing missing values (geom_point).

Inference?

Correlation Analysis

```
# precej zanimivi so Height, Length, Numinversions
(cor.test(roller_coasters_raw$Height, roller_coasters_raw$Speed))
```

```
##
##  Pearson's product-moment correlation
##
## data:  roller_coasters_raw$Height and roller_coasters_raw$Speed
## t = 38.222, df = 256, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9019179 0.9388051
## sample estimates:
##       cor
## 0.9224392
```

```
(cor.test(roller_coasters_raw$Length, roller_coasters_raw$Speed))
```

```
##
```

```
##  Pearson's product-moment correlation
##
## data:  roller_coasters_raw$Length and roller_coasters_raw$Speed
## t = 15.582, df = 258, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6278199 0.7540719
## sample estimates:
##       cor
## 0.6962931
```

```
(cor.test(roller_coasters_raw$Numinversions, roller_coasters_raw$Speed))
```

```
##
##  Pearson's product-moment correlation
##
## data:  roller_coasters_raw$Numinversions and roller_coasters_raw$Speed
## t = 5.5742, df = 268, p-value = 6.061e-08
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2110692 0.4253337
## sample estimates:
##       cor
## 0.3223236
```

```
(cor.test(roller_coasters_raw$Duration, roller_coasters_raw$Speed))
```

```
##
##  Pearson's product-moment correlation
##
## data:  roller_coasters_raw$Duration and roller_coasters_raw$Speed
## t = 3.9954, df = 162, p-value = 9.781e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1532868 0.4328823
## sample estimates:
##       cor
## 0.2995011
```

```
(cor.test(roller_coasters_raw$GForce, roller_coasters_raw$Speed))
```

```
##
##  Pearson's product-moment correlation
##
## data:  roller_coasters_raw$GForce and roller_coasters_raw$Speed
## t = 3.3676, df = 56, p-value = 0.001377
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1701111 0.6045861
## sample estimates:
##       cor
## 0.4103754
```

```
(cor.test(roller_coasters_raw$Opened, roller_coasters_raw$Speed)) # shit
```

```
##
##  Pearson's product-moment correlation
##
## data:  roller_coasters_raw$Opened and roller_coasters_raw$Speed
## t = 0.26238, df = 260, p-value = 0.7932
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1051251  0.1371870
## sample estimates:
##        cor
## 0.01626982
```

Regression

```
roller_coasters <- roller_coasters_raw %>%
  select(Construction, Length, Height, Speed) %>%
  filter(!is.na(Speed) & !is.na(Height) & !is.na(Length)) %>%
  mutate("Steel" = as.numeric(Construction == 'Steel')) %>%
  select(-Construction)
roller_coasters
```

```
## # A tibble: 252 x 4
##     Length Height Speed Steel
##      <dbl>  <dbl> <dbl> <dbl>
## 1    853.  128.   194.     1
## 2    376.  126.   162      1
## 3   2010.   94.5  151.     1
## 4   1619.   74.7  138.     1
## 5   1620    73    127      1
## 6   1372.   71.6  138.     1
## 7   1610.   70.1  130.     1
## 8   1700    67    120      1
## 9   2143.   66.4  126.     0
## 10   396.   66.4  113.     1
## # ... with 242 more rows
```

```
## 75% of the sample size
smp_size <- floor(0.75 * nrow(roller_coasters))

## set the seed to make your partition reproducible
set.seed(123)
train_ind <- sample(seq_len(nrow(roller_coasters)), size = smp_size)

(train <- roller_coasters[train_ind, ])
```

```
## # A tibble: 189 x 4
##     Length Height Speed Steel
##      <dbl>  <dbl> <dbl> <dbl>
```

```
## 1    412.    16.2  50.2      1
## 2    207      8.5  34.9      1
## 3    538.    13.4  50.2      1
## 4    375.    61    105.      1
## 5    427.    10.7  32.4      0
## 6    774.    14.6  68.0      1
## 7    950     36    90        1
## 8    717.    23.8  82.6      1
## 9    309.    39.9  81        1
## 10   264      6    45        1
## # ... with 179 more rows
```

```
(test <- roller_coasters[-train_ind, ])
```

```
## # A tibble: 63 x 4
##     Length Height Speed Steel
##      <dbl>  <dbl> <dbl> <dbl>
## 1    376.  126.    162      1
## 2   2010.   94.5  151.      1
## 3    671.   62.5  133.      1
## 4    347.   61    105.      1
## 5    367.   58.2  105.      1
## 6   1167.   57.3  113.      1
## 7   1654.   49.1  105.      0
## 8   1332.   47.5  105.      1
## 9    150    46    105       1
## 10   192.   45.7  105.      1
## # ... with 53 more rows
```

```
lin_model <- lm(Speed ~ ., data = train)
(summary(lin_model))
```

```
##
## Call:
## lm(formula = Speed ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.388  -6.020  -0.644   4.466  35.365
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.376471   2.513555  13.279  < 2e-16 ***
## Length       0.013174   0.002178   6.047 7.97e-09 ***
## Height       1.248969   0.047545  26.269  < 2e-16 ***
## Steel       -5.752860   2.109654  -2.727  0.00701 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.749 on 185 degrees of freedom
## Multiple R-squared:  0.9103, Adjusted R-squared:  0.9088
## F-statistic: 625.7 on 3 and 185 DF,  p-value: < 2.2e-16
```

14

```
(coef(lin_model))
```

```
## (Intercept)      Length      Height      Steel
## 33.37647051  0.01317372  1.24896948 -5.75286049
```

Second Dataset Analysis –> TODO

State hypothesis here...