R Notebook

# TidyTuesday

Join the R4DS Online Learning Community in the weekly #TidyTuesday event! Every week we post a raw dataset, a chart or article related to that dataset, and ask you to explore the data. While the dataset will be “tamed”, it will not always be tidy! As such you might need to apply various R for Data Science techniques to wrangle the data into a true tidy format. The goal of TidyTuesday is to apply your R skills, get feedback, explore other’s work, and connect with the greater #RStats community! As such we encourage everyone of all skills to participate!

# Load the weekly Data

Download the weekly data and make available in the tt object.

tt <- tt\_load("2021-05-25")

## --- Compiling #TidyTuesday Information for 2021-05-25 ----

## --- There are 2 files available ---

## --- Starting Download ---

##   
## Downloading file 1 of 2: `drivers.csv`  
## Downloading file 2 of 2: `records.csv`

## --- Download complete ---

drivers <- tt$drivers  
records <- tt$records  
rm(tt)

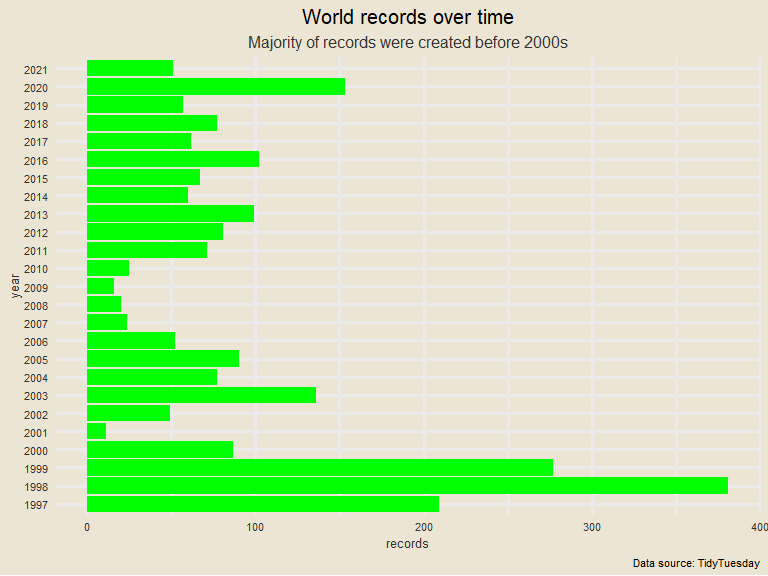
# Wrangle

Explore the data and process it into a nice format for plotting! Access each dataset by name by using a dollarsign after the tt object and then the name of the data set.

# Visualize

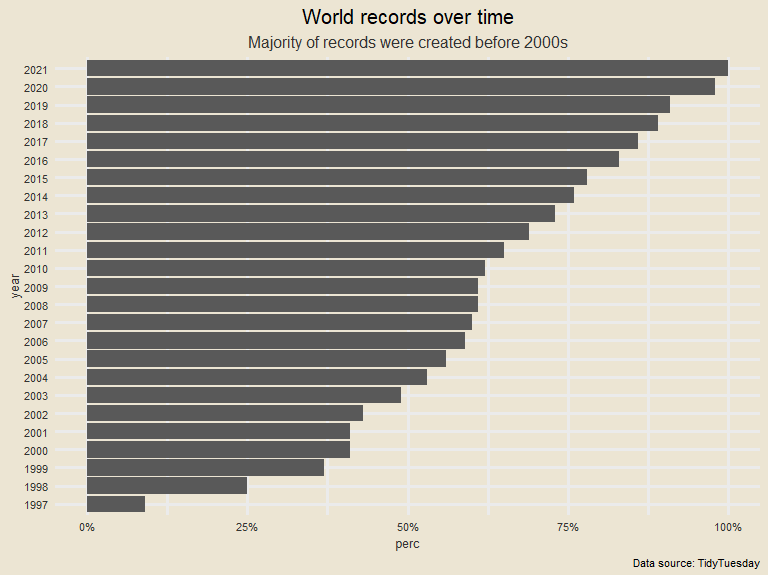
## World records over time

records %>%   
 mutate(year=format(date,"%Y")) %>%   
 group\_by(year) %>%   
 summarise(records = n()) %>%   
 ggplot(aes(year,records))+  
 geom\_bar(stat = "identity",fill="green")+  
 labs(title = "World records over time",  
 subtitle = str\_wrap("Majority of records were created before 2000s"),  
 caption = "Data source: TidyTuesday")+  
 theme(plot.title = element\_text(hjust=0.5,size = 15),  
 plot.subtitle = element\_text(hjust=0.5))+  
 coord\_flip()



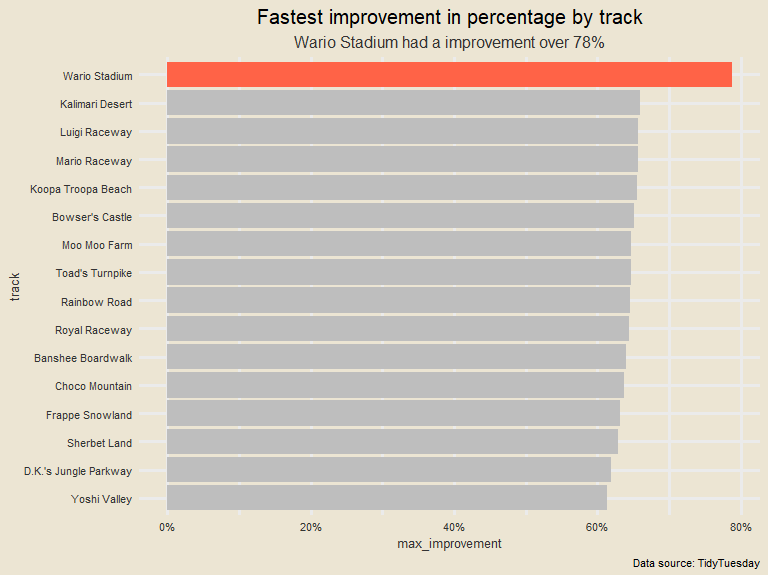
## World records over time in cumulative percentage

records %>%   
 mutate(year=format(date,"%Y")) %>%   
 group\_by(year) %>%   
 summarise(records=n()) %>%   
 mutate(cum\_perc = cumsum(records),  
 perc=round(cum\_perc/sum(records),2)) %>%   
 ggplot(aes(year,perc))+  
 geom\_bar(stat = "identity")+  
 labs(title = "World records over time",  
 subtitle = str\_wrap("Majority of records were created before 2000s"),  
 caption = "Data source: TidyTuesday")+  
 theme(plot.title = element\_text(hjust=0.5,size = 15),  
 plot.subtitle = element\_text(hjust=0.5))+  
 coord\_flip()+  
 scale\_y\_continuous(labels = percent\_format())



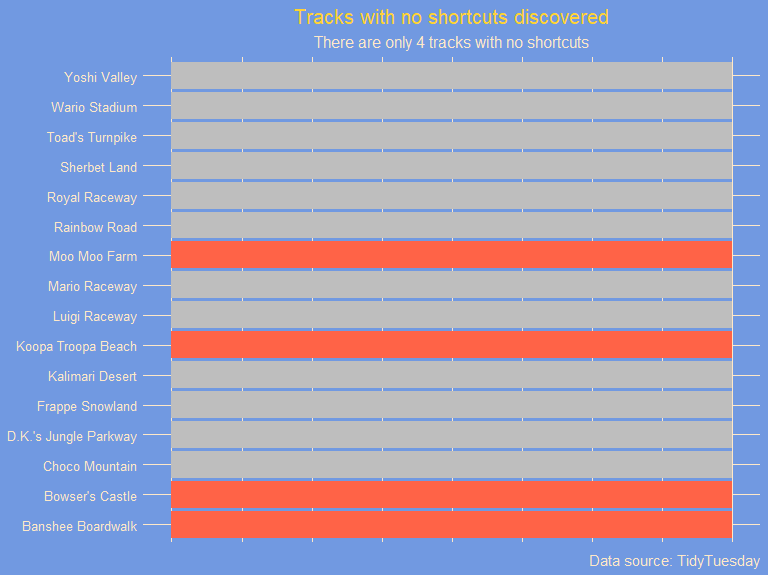
## For which track did the world record improve the most?

improvement\_by\_track <- records %>%   
 group\_by(track) %>%   
 mutate(improvement = 1-time/lag(time)) %>%   
 summarise(max\_improvement = max(improvement,na.rm = TRUE)) %>%   
 mutate(track = fct\_reorder(track,max\_improvement)) %>%   
 arrange(desc(max\_improvement)) %>%   
 mutate(ToHighlight =ifelse(max\_improvement == max(max\_improvement), "yes", "no"))  
   
 ggplot(improvement\_by\_track,aes(track,max\_improvement,fill=ToHighlight ))+   
 geom\_bar(stat = "identity")+  
 coord\_flip()+  
 scale\_y\_continuous(labels = percent\_format())+  
 labs(title = "Fastest improvement in percentage by track",  
 subtitle = str\_wrap("Wario Stadium had a improvement over 78%"),  
 caption = "Data source: TidyTuesday")+  
 theme(plot.title = element\_text(hjust=0.5,size = 15),  
 plot.subtitle = element\_text(hjust=0.5))+  
 scale\_fill\_manual( values = c( "yes"="tomato", "no"="gray" ), guide = FALSE )



## For how many tracks have shortcuts been discovered?

records %>%   
 group\_by(track,shortcut) %>%  
 count() %>%   
 spread(shortcut,n) %>%   
 mutate(shortcuts\_discovered = ifelse(Yes<1 | is.na(Yes),"No","Yes")) %>%   
 select(-No,-Yes) %>%   
 ggplot(aes(track,fill=shortcuts\_discovered))+   
 geom\_bar()+  
 coord\_flip()+  
 labs(title = "Tracks with no shortcuts discovered",  
 subtitle = str\_wrap("There are only 4 tracks with no shortcuts"),  
 caption = "Data source: TidyTuesday")+  
 theme\_simpsons()+  
 theme(plot.title = element\_text(hjust=0.5,size = 15),  
 plot.subtitle = element\_text(hjust=0.5),  
 axis.title = element\_blank(),  
 axis.text.x = element\_blank())+  
 scale\_fill\_manual( values = c( "Yes"="gray", "No"="tomato" ), guide = FALSE )



## When were shortcuts discovered?

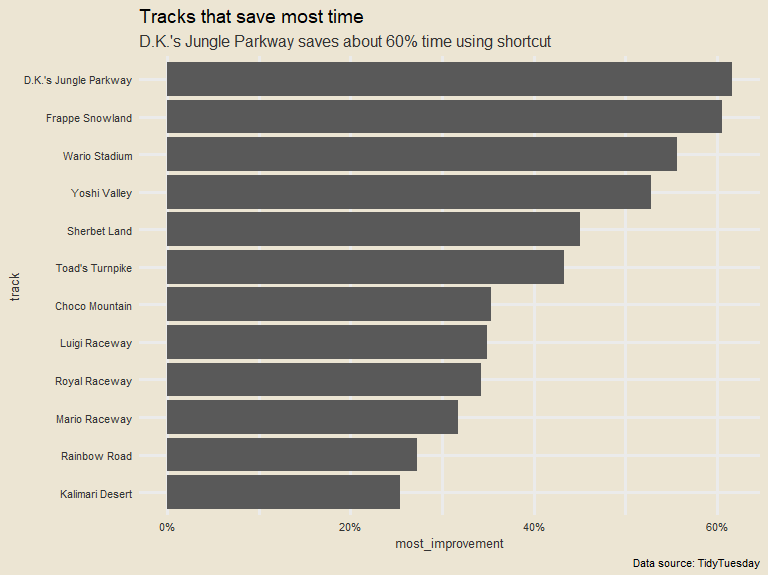
records %>%   
 group\_by(track) %>%   
 filter(shortcut=="Yes") %>%   
 summarise(min\_date = min(date))

## # A tibble: 12 x 2  
## track min\_date   
## <chr> <date>   
## 1 Choco Mountain 1997-03-10  
## 2 D.K.'s Jungle Parkway 1997-03-10  
## 3 Frappe Snowland 1997-03-10  
## 4 Kalimari Desert 1997-03-10  
## 5 Luigi Raceway 1997-02-16  
## 6 Mario Raceway 1997-03-10  
## 7 Rainbow Road 1997-03-07  
## 8 Royal Raceway 1997-03-10  
## 9 Sherbet Land 1997-03-10  
## 10 Toad's Turnpike 1997-03-10  
## 11 Wario Stadium 1997-03-10  
## 12 Yoshi Valley 1997-03-07

## On which track does the shortcut save the most time?

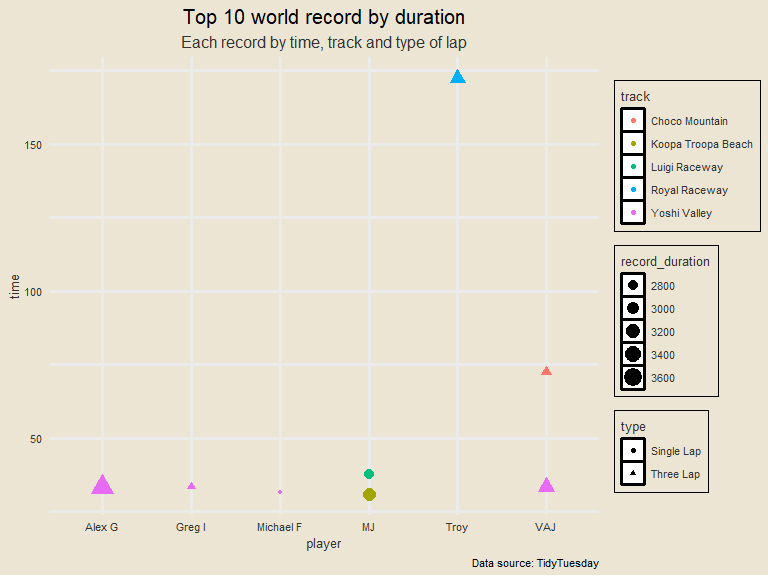
records %>%   
 group\_by(track,shortcut) %>%   
 summarise(ave\_time= mean(time,na.rm = TRUE)) %>%   
 spread(shortcut,ave\_time) %>%   
 ungroup() %>%   
 mutate(most\_improvement = 1-Yes/No) %>%   
 mutate(track = fct\_reorder(track,most\_improvement)) %>%   
 arrange(most\_improvement) %>%   
 filter(!is.na(most\_improvement)) %>%   
 ggplot(aes(track,most\_improvement))+  
 geom\_bar(stat = "identity")+  
 coord\_flip()+  
 scale\_y\_continuous(labels = percent\_format())+  
 labs(title = "Tracks that save most time",  
 subtitle = str\_wrap("D.K.'s Jungle Parkway saves about 60% time using shortcut"),  
 caption = "Data source: TidyTuesday")

## `summarise()` has grouped output by 'track'. You can override using the `.groups` argument.



## Which is the longest standing world record?

records %>%   
 arrange(desc(record\_duration)) %>%   
 slice(1:10) %>%   
 ggplot(aes(player,time,color=track,size=record\_duration,shape=type))+  
 geom\_point()+  
 labs(title = "Top 10 world record by duration",  
 subtitle = str\_wrap("Each record by time, track and type of lap"),  
 caption = "Data source: TidyTuesday")+  
 theme(plot.title = element\_text(hjust=0.5,size = 15),  
 plot.subtitle = element\_text(hjust=0.5),  
 legend.position = "right")



## Who is the player with the most world records?

drivers %>%   
 group\_by(player,nation) %>%   
 summarise(records = sum(records,na.rm = TRUE)) %>%   
 ungroup() %>%   
 mutate(player = fct\_reorder(player,records)) %>%   
 arrange(desc(records)) %>%   
 slice(1:10) %>%   
 ggplot(aes(player,records,fill=nation))+  
 geom\_bar(stat = "identity")+  
 coord\_flip()+  
 labs(title = "Top 10 players with world records",  
 caption = "Data source: TidyTuesday")+  
 theme(plot.title = element\_text(hjust=0.5,size = 15),  
 plot.subtitle = element\_text(hjust=0.5),  
 legend.position = "right")+  
 geom\_label(aes(label=records),show.legend = FALSE,hjust=1,col="white",size=3)

## `summarise()` has grouped output by 'player'. You can override using the `.groups` argument.

