Time Built P5

Chloe Hall

2022-11-16

```
#Load necessary packages
library(readxl)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(dbplyr)
```

```
##
## Attaching package: 'dbplyr'
##
## The following objects are masked from 'package:dplyr':
##
##     ident, sql
```

```
#Load the data
enr <- read_excel("~/Downloads/ROCCA/ENR facility spreadsheet april.xlsx")

#select to relevant columns
enr<-enr %>%
  select(country_name, ccode, facility_name, construction_start, construction_start_lower_bound, constru
```

```
#list of countries in dataset
country<-unique(enr$country_name)
```

```
final <- data.frame(matrix(ncol = 6, nrow = 0))
colnames(final)<- c("country_name", "ccode", "facility_name", "start", "end")

for (row in 1:nrow(enr)) {
  if (is.na(enr[row, 4]) || enr[row, 4] < 0||enr[row, 4] > 3000) {
    start = enr$construction_start_lower_bound[row]
  }
  else{
    start = enr$construction_start[row]
  }
```

```
  if (is.na(enr[row, 7] )|| enr[row, 7] < 0||enr[row, 7] > 3000) {
     end= enr$construction_end_lower_bound[row]
  }  else{
    end = enr$construction_end[row]
  }


  if(!is.na(start)&!is.na(end)){
      final[nrow(final) + 1, ] <-
      c(enr[row, 1], enr[row, 2], enr[row, 3], start, end, enr[row,10])
  }
}
```

```
df<- data.frame(matrix(ncol = 5, nrow = 0))
colnames(df)<- c("country_name", "ccode", "start", "years_to_build", "enr_type")

#Creating Variable for build time
for (row in 1:nrow(final)) {
  years_to_build = final$end[row]-final$start[row]

    if(years_to_build>=0){ #Removing the weird negative range values
      df[nrow(df) + 1, ] <-
      c(final[row, 1], final[row, 2], final[row, 4], years_to_build, final[row,6])
    }
}
```

```
str(df)
```

```
## 'data.frame':    240 obs. of  5 variables:
##  $ country_name  : chr  "Algeria" "Argentina" "Argentina" "Argentina" ...
##  $ ccode         : chr  "615" "160" "160" "160" ...
##  $ start         : chr  "1986" "1968" "1978" "1979" ...
##  $ years_to_build: chr  "6" "0" "12" "8" ...
##  $ enr_type      : chr  "1" "1" "1" "2" ...
```

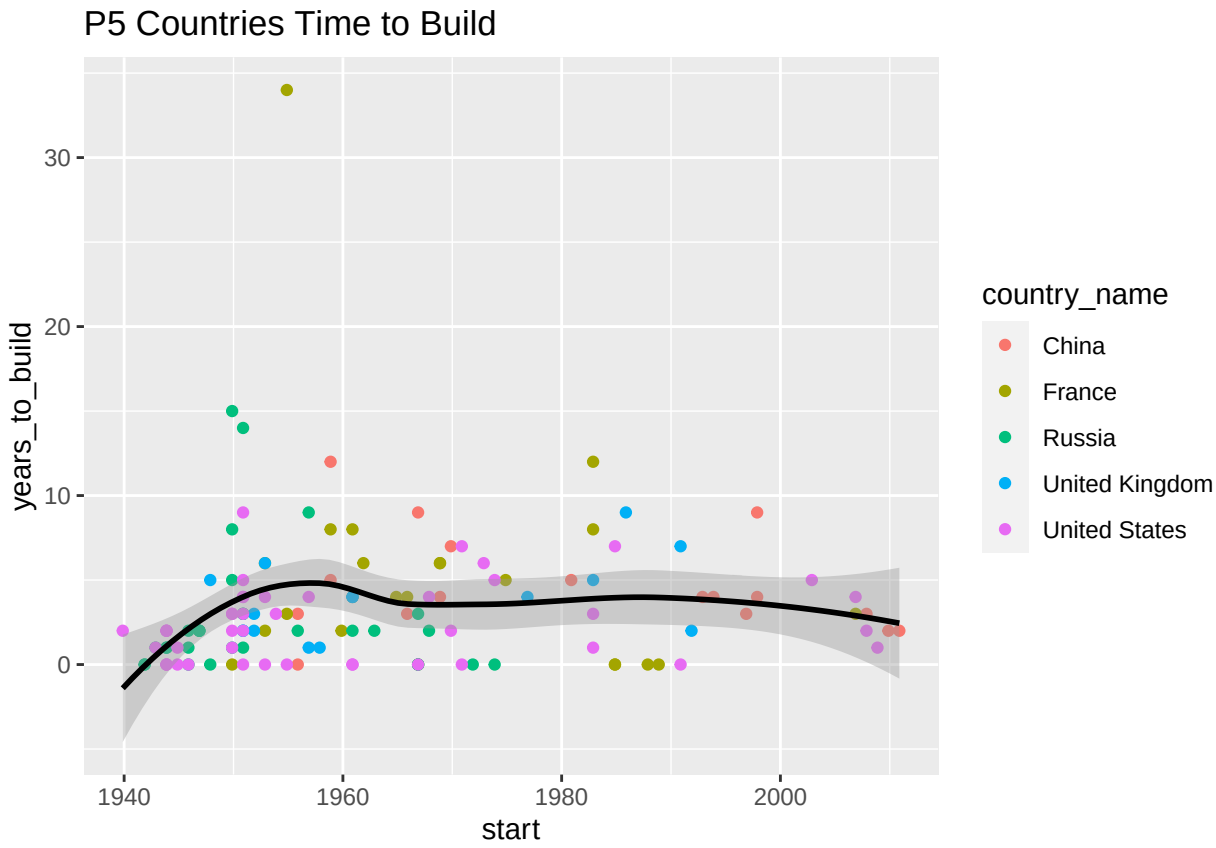
```
df$start<-as.Date(as.character(df$start), format = "%Y")
df$years_to_build<-as.numeric(df$years_to_build)
str(df)
```

```
## 'data.frame':    240 obs. of  5 variables:
##  $ country_name  : chr  "Algeria" "Argentina" "Argentina" "Argentina" ...
##  $ ccode         : chr  "615" "160" "160" "160" ...
##  $ start         : Date, format: "1986-11-16" "1968-11-16" ...
##  $ years_to_build: num  6 0 12 8 7 10 6 6 11 12 ...
##  $ enr_type      : chr  "1" "1" "1" "2" ...
```

```
p5<-df %>%
  filter(country_name=="China"|country_name=="France"|country_name=="Russia"|country_name=="United King
```

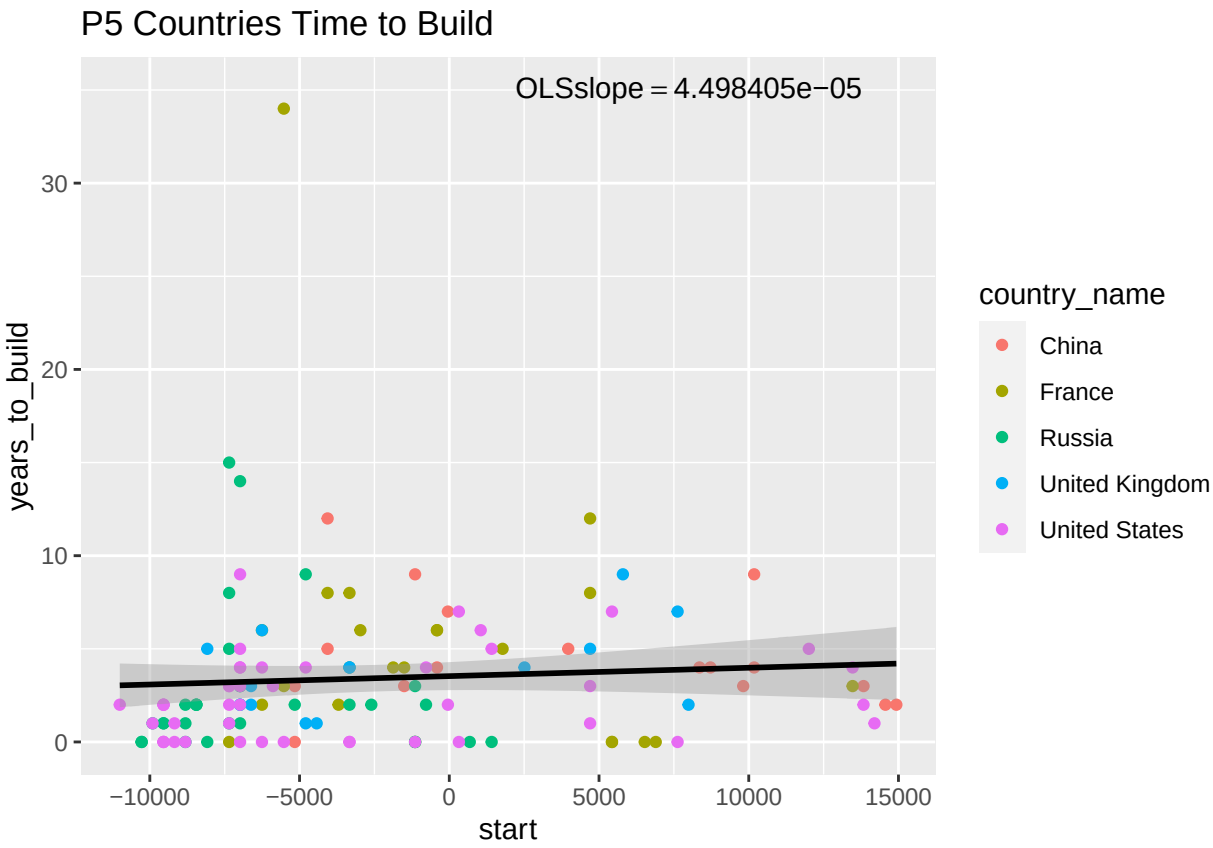
```
ggplot(p5, aes(x=start, y=years_to_build, color=country_name)) +
  geom_point()+
  geom_smooth(method = loess, se = T, color = "black")+
  ggtitle("P5 Countries Time to Build")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
ggplot(p5, aes(x=start, y=years_to_build, color=country_name)) +
  annotate("text", x=8000,y=35,label=(paste0("OLSslope==",coef(lm(p5$years_to_build~p5$start))[2])),pars
  geom_point()+
  geom_smooth(method = lm, se = T, color = "black")+
  ggtitle("P5 Countries Time to Build")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



With No Outliers

```
#Calculating the outliers
Q <- quantile(df$years_to_build, probs=c(.25, .75), na.rm = FALSE)

iqr <- IQR(df$years_to_build)

up <-  Q[2]+1.5*iqr # Upper Range
up
```

```
##  75%
## 13.5
```

```
low<- Q[1]-1.5*iqr # Lower Range
low
```

```
##  25%
## -6.5
```

```
outliers<- subset(df, df$years_to_build < (Q[1] - 1.5*iqr) | df$years_to_build > (Q[2]+1.5*iqr))
outliers
```

```
##       country_name ccode      start years_to_build enr_type
## 13          Brazil   140 1960-11-16             22        1
## 35  Czech Republic   315 1955-11-16             22        1
## 42          France   220 1954-11-16             34        1
## 83            Iran   630 1987-11-16             18        1
## 95          Israel   666 1958-11-16             21        3
## 98          Israel   666 1964-11-16             19        1
## 126       Pakistan   770 1974-11-16             41        1
```

```
## 137          Russia   365 1949-11-16              15         2
## 139          Russia   365 1950-11-16              14         1
```

```
nonoutliers<-subset(df, df$years_to_build > (Q[1] - 1.5*iqr) & df$years_to_build < (Q[2]+1.5*iqr))
```

```
p5norm<-nonoutliers %>%
  filter(country_name=="China"|country_name=="France"|country_name=="Russia"|country_name=="United King
```

```
ggplot(p5norm, aes(x=start, y=years_to_build, color=country_name)) +
  geom_point()+
  geom_smooth(method = loess, se = T, color = "black")+
  ggtitle("P5 Countries Time to Build")
```

```
## `geom_smooth()` using formula 'y ~ x'
```