transit costs 1-05

Amber Lee

1/5/2021

```
library(tidytuesdayR)
library(tidyverse)
library(countrycode)
```

```
tuesdata <- tidytuesdayR::tt_load('2021-01-05')
```

```
##
##  Downloading file 1 of 1: `transit_cost.csv`
```

```
transit_cost <- tuesdata$transit_cost
```

exploration

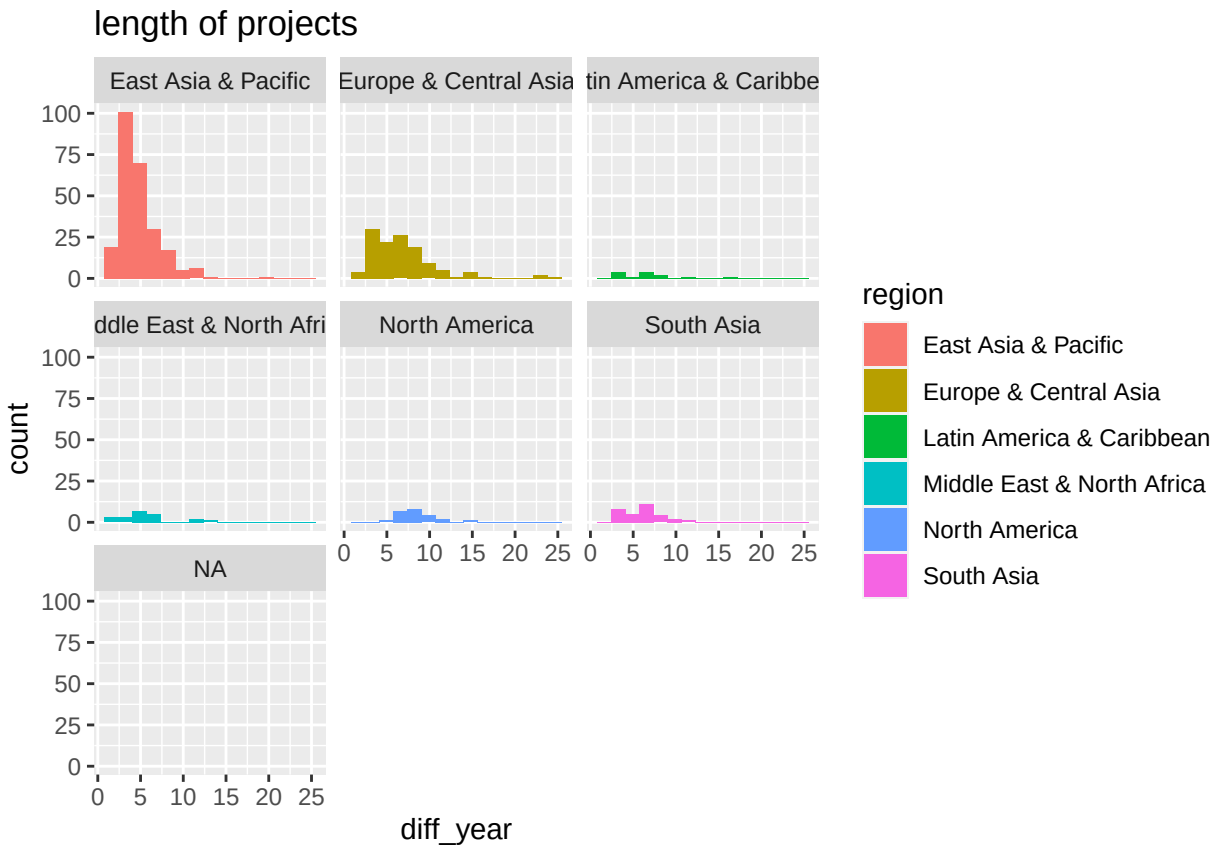
```
# trying out countrycode package

transit_cost <- transit_cost %>%
  mutate(region = countrycode(country, origin = "ecb",
                                    destination = "region")) %>%
  mutate(region = case_when(country == "UK" ~ "Europe & Central Asia",
                            TRUE ~ region))
```

variables to make numeric: start_year, end_year real_cost tunnel_per

```
transit_cost <- transit_cost %>%
  mutate(start_year = as.numeric(start_year),
         end_year = as.numeric(end_year),
         diff_year = end_year - start_year,
         real_cost = as.numeric(real_cost),
         tunnel_per = as.numeric(str_remove(tunnel_per, "%")))
```

```
ggplot(data = transit_cost, aes(x = diff_year, fill = region)) +
  geom_histogram(bins = 15) +
  facet_wrap(~ region) +
  labs(title = "length of projects")
```



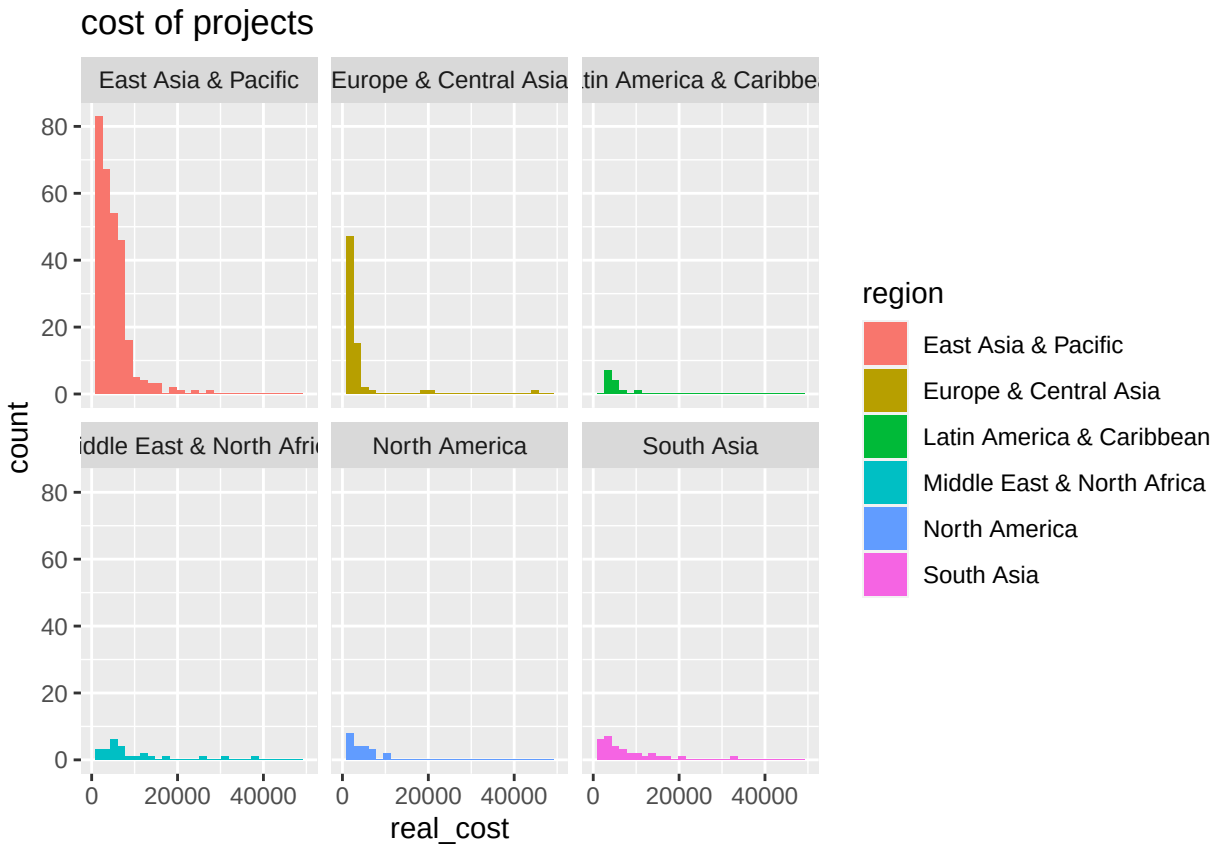

```
transit_cost %>%
  filter(is.na(diff_year)) %>%
  select(start_year, end_year, diff_year)
```

```
## # A tibble: 82 x 3
##    start_year end_year diff_year
##         <dbl>    <dbl>     <dbl>
##  1       2021       NA        NA
##  2         NA     2020        NA
##  3         NA     2019        NA
##  4       2020       NA        NA
##  5       2020       NA        NA
##  6         NA       NA        NA
##  7       2020       NA        NA
##  8       2019       NA        NA
##  9         NA       NA        NA
## 10         NA       NA        NA
## # ... with 72 more rows
```

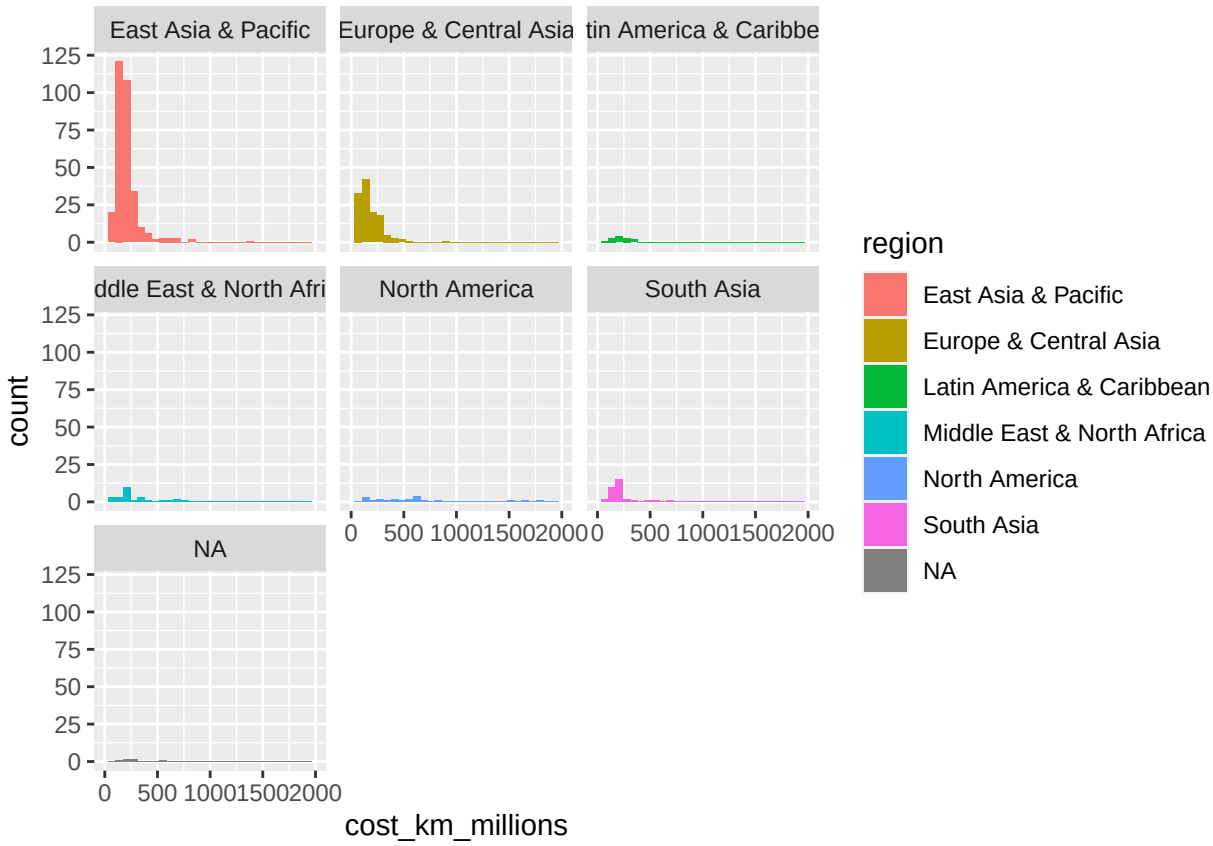
```
# NAs are from projects that are still in construction or start year not known
```

```
transit_cost %>%
  filter(!is.na(region)) %>%
  ggplot(aes(x = real_cost, fill = region)) +
  geom_histogram() +
  facet_wrap(~ region) +
  labs(title = "cost of projects") +
  # scale to see distribution better, about a dozen observations excluded
```

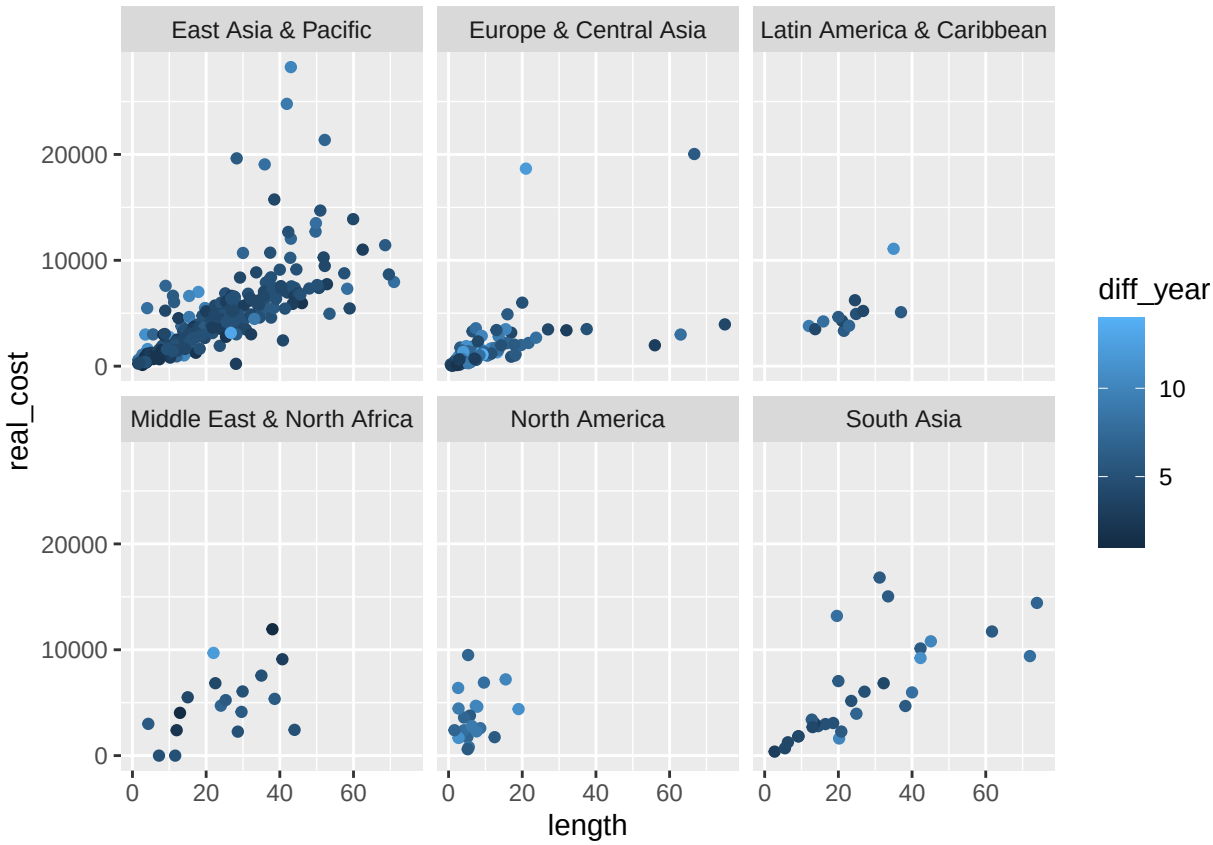
```
  scale_x_continuous(limits = c(0, 50000),
                     breaks = c(0, 20000, 40000))
```



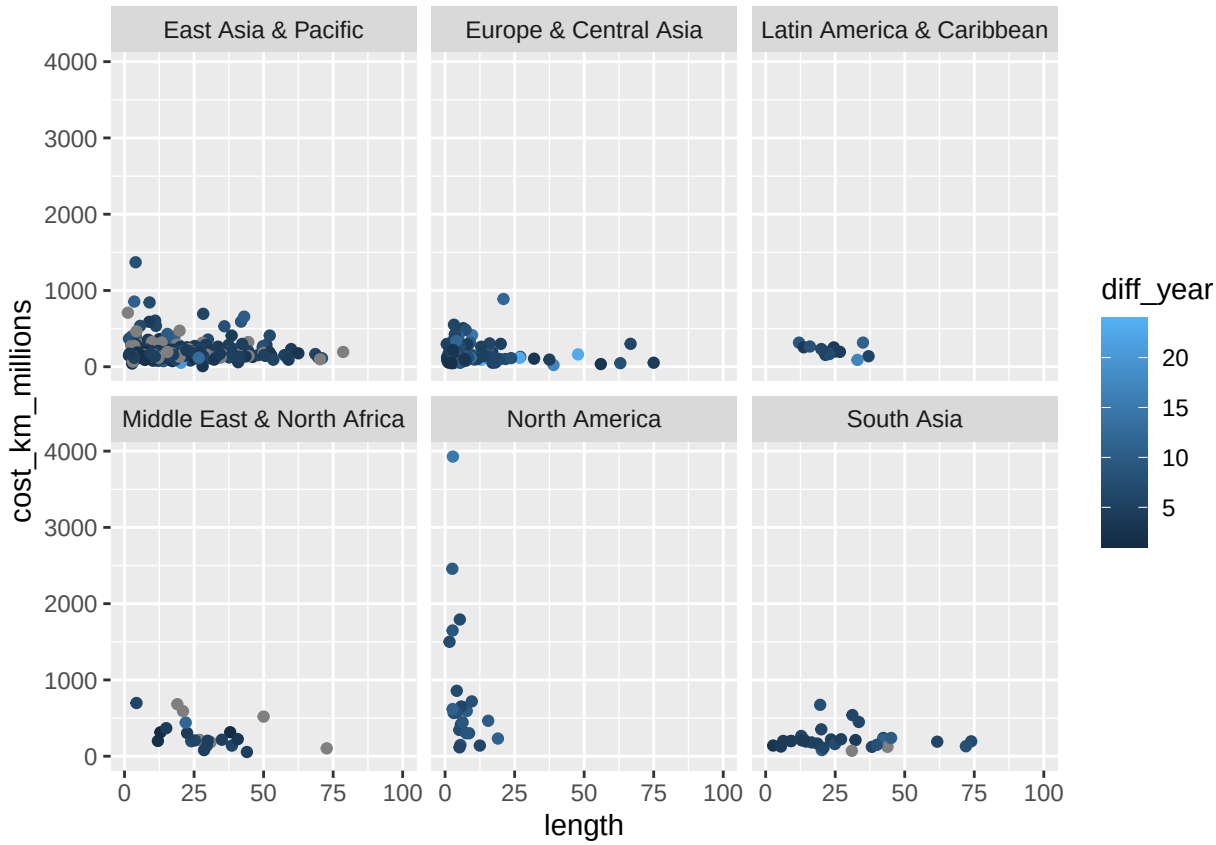

```
ggplot(transit_cost, aes(x = cost_km_millions, fill = region)) +
  geom_histogram() +
  facet_wrap(~ region) +
  # scale to see distribution better, about a dozen observations excluded
  scale_x_continuous(limits = c(0, 2000))
```




```
transit_cost %>%
  filter(!is.na(country) & diff_year < 15 &
           real_cost < 50000 & length < 100) %>%
  ggplot(aes(x = length, y = real_cost)) +
  geom_point(aes(color = diff_year)) +
  facet_wrap(~ region)
```

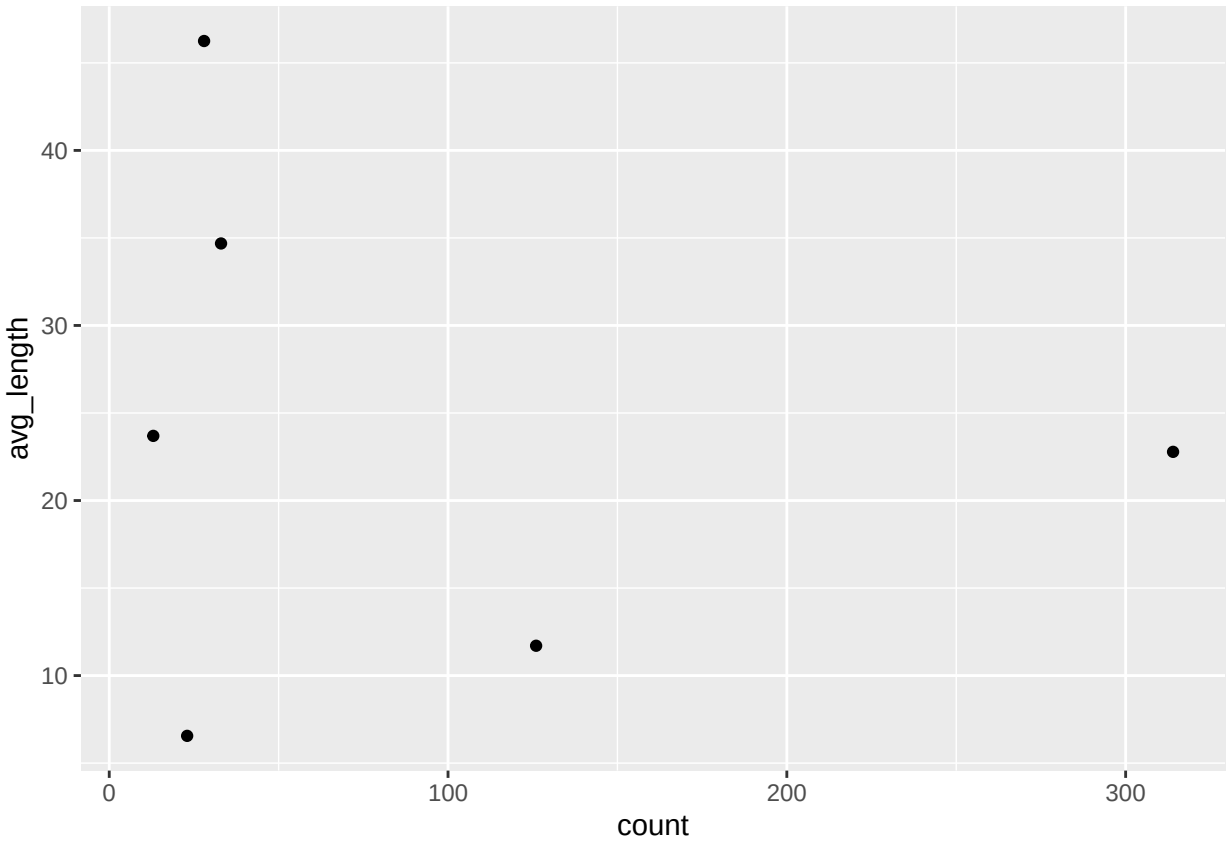



```
transit_cost %>%
  filter(!is.na(country)) %>%
  ggplot(aes(x = length, y = cost_km_millions)) +
  geom_point(aes(color = diff_year)) +
  facet_wrap(~ region) +
  scale_x_continuous(limits = c(0, 100))
```




```
reg_transit_cost <- transit_cost %>%
  group_by(region, city) %>%
  summarize(count = n(),
            avg_length = mean(length),
            avg_real_cost = mean(real_cost),
            avg_kmpermil = mean(cost_km_millions),
            avg_diff = mean(diff_year))

transit_cost %>%
  group_by(region) %>%
  summarize(count = n(),
            avg_length = mean(length),
            avg_real_cost = mean(real_cost),
            avg_kmpermil = mean(cost_km_millions),
            avg_diff = mean(diff_year)) %>%
  ggplot(aes(x = count, y = avg_length)) +
  geom_point()
```


avg_length
40
30
20
10
0
100
200
300
count