Econ 216 Final Project

Kyler Rosen, Raunak Bhumsaria, Gordon Bradley, and Alex Illchev

2024-03-31

Load and Inspect Data

```
# Load necessary libraries
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)
library(ggmap)
```

```
## Warning: package 'ggmap' was built under R version 4.3.1
```

```
## i Google's Terms of Service: <https://mapsplatform.google.com>
##   Stadia Maps' Terms of Service: <https://stadiamaps.com/terms-of-service/>
##   OpenStreetMap's Tile Usage Policy: <https://operations.osmfoundation.org/policies/tiles/>
## i Please cite ggmap if you use it! Use `citation("ggmap")` for details.
```

```
us_map <- map_data("state")


# Load the data
data <- read.csv("uswtdb_v4_3_20220114.csv")

# Inspect the first few rows of the data
head(data)
```

```
##   case_id   faa_ors          faa_asn usgs_pr_id eia_id t_state     t_county
## 1 3072661                                  5149  52161      CA  Kern County
## 2 3072695                                  5143  52161      CA  Kern County
## 3 3072704                                  5146  52161      CA  Kern County
## 4 3063272 19-028134 2014-WTE-4084-OE         NA     NA      IA Story County
## 5 3053390 19-028015 2015-WTE-6386-OE         NA     NA      IA Boone County
## 6 3063269 19-028130 2016-WTE-5934-OE         NA     NA      IA Story County
##   t_fips                  p_name p_year p_tnum p_cap t_manu    t_model t_cap
## 1   6029                251 Wind   1987    194 18.43 Vestas               95
## 2   6029                251 Wind   1987    194 18.43 Vestas               95
## 3   6029                251 Wind   1987    194 18.43 Vestas               95
## 4  19169 30 MW Iowa DG Portfolio   2017     10 30.00 Nordex AW125/3000  3000
## 5  19015 30 MW Iowa DG Portfolio   2017     10 30.00 Nordex AW125/3000  3000
## 6  19169 30 MW Iowa DG Portfolio   2017     10 30.00 Nordex AW125/3000  3000
##   t_hh t_rd    t_rsa t_ttlh retrofit retrofit_year t_conf_atr t_conf_loc
## 1   NA   NA       NA     NA        0            NA          2          3
## 2   NA   NA       NA     NA        0            NA          2          3
## 3   NA   NA       NA     NA        0            NA          2          3
## 4 87.5  125 12271.85    150        0            NA          3          3
## 5 87.5  125 12271.85    150        0            NA          3          3
## 6 87.5  125 12271.85    150        0            NA          3          3
##   t_img_date    t_img_srce      xlong     ylat
## 1   5/8/2018 Digital Globe -118.36376 35.07791
## 2   5/8/2018 Digital Globe -118.36441 35.07744
## 3   5/8/2018 Digital Globe -118.36420 35.07764
## 4  4/24/2017 Digital Globe  -93.43037 42.02823
## 5   6/1/2017 Digital Globe  -93.70042 41.97761
## 6  7/23/2017 Digital Globe  -93.63284 41.88248
```

```
# Summarize the data to understand its structure
summary(data)
```

```
##     case_id          faa_ors            faa_asn            usgs_pr_id
##  Min.   :3000001   Length:70808       Length:70808       Min.   :    1
##  1st Qu.:3032230   Class :character   Class :character   1st Qu.:18626
##  Median :3050978   Mode  :character   Mode  :character   Median :28598
##  Mean   :3058490                                         Mean   :27524
##  3rd Qu.:3090448                                         3rd Qu.:38720
##  Max.   :3118671                                         Max.   :49135
##                                                          NA's   :32545
##      eia_id        t_state            t_county             t_fips
##  Min.   :   90   Length:70808       Length:70808       Min.   : 2013
##  1st Qu.:56763   Class :character   Class :character   1st Qu.:19081
##  Median :57752   Mode  :character   Mode  :character   Median :35057
##  Mean   :57878                                         Mean   :32244
##  3rd Qu.:60338                                         3rd Qu.:48141
##  Max.   :65270                                         Max.   :72133
##  NA's   :5793
##     p_name              p_year        p_tnum           p_cap
##  Length:70808       Min.   :1981   Min.   :  1.0   Min.   :  0.05
##  Class :character   1st Qu.:2008   1st Qu.: 56.0   1st Qu.: 99.00
##  Mode  :character   Median :2012   Median : 85.0   Median :158.00
##                     Mean   :2012   Mean   :104.4   Mean   :170.18
##                     3rd Qu.:2018   3rd Qu.:121.0   3rd Qu.:211.22
##                     Max.   :2021   Max.   :731.0   Max.   :525.02
##                     NA's   :613                    NA's   :4482
##     t_manu            t_model              t_cap            t_hh
##  Length:70808       Length:70808       Min.   :  50   Min.   : 19.00
##  Class :character   Class :character   1st Qu.:1500   1st Qu.: 80.00
##  Mode  :character   Mode  :character   Median :2000   Median : 80.00
##                                        Mean   :1964   Mean   : 81.06
##                                        3rd Qu.:2300   3rd Qu.: 87.00
##                                        Max.   :6000   Max.   :131.00
##                                        NA’s   :5480   NA’s   :6180
##       t_rd           t_rsa           t_ttlh         retrofit
##  Min.   : 13.40   Min.   :  141   Min.   : 30.4   Min.   :0.00000
##  1st Qu.: 82.00   1st Qu.: 5281   1st Qu.:121.0   1st Qu.:0.00000
##  Median :100.00   Median : 7854   Median :130.1   Median :0.00000
##  Mean   : 95.66   Mean   : 7619   Mean   :129.1   Mean   :0.08454
##  3rd Qu.:110.00   3rd Qu.: 9503   3rd Qu.:145.1   3rd Qu.:0.00000
##  Max.   :155.00   Max.   :18869   Max.   :199.6   Max.   :1.00000
##  NA’s   :5934     NA’s   :5934    NA’s   :6180
##  retrofit_year     t_conf_atr      t_conf_loc     t_img_date
##  Min.   :2015    Min.   :1.000   Min.   :1.000   Length:70808
##  1st Qu.:2018    1st Qu.:3.000   1st Qu.:3.000   Class :character
##  Median :2019    Median :3.000   Median :3.000   Mode  :character
##  Mean   :2019    Mean   :2.767   Mean   :2.884
##  3rd Qu.:2020    3rd Qu.:3.000   3rd Qu.:3.000
##  Max.   :2020    Max.   :3.000   Max.   :3.000
##  NA’s   :64822
##   t_img_srce            xlong              ylat
##  Length:70808       Min.   :-171.71   Min.   :13.39
##  Class :character   1st Qu.:-103.04   1st Qu.:34.43
##  Mode  :character   Median : -99.39   Median :39.05
##                     Mean   :-100.09   Mean   :38.48
##                     3rd Qu.: -95.20   3rd Qu.:42.81
##                     Max.   : 144.72   Max.   :66.84
##
```

Exploring Relationships

Analyze relationships between variables using scatter plots, box plots, etc.

Scatter plot for two continuous variables

ggplot(data, aes(x = continuous_variable_1, y = continuous_variable_2)) + geom_point() + labs(title = “Scatter Plot of Variable 1 vs Variable 2”, x = “Variable 1”, y = “Variable 2”)

Box plot for a continuous and a categorical variable

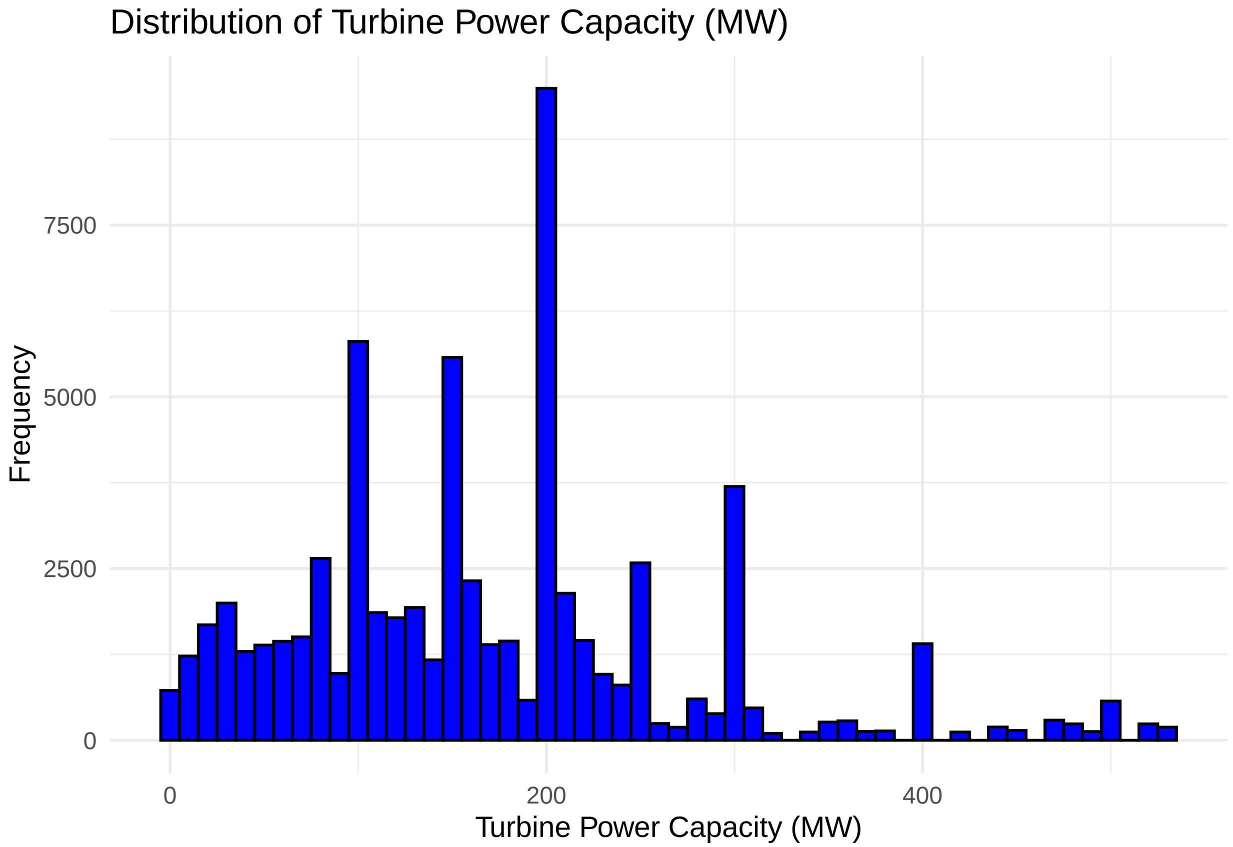
ggplot(data, aes(x = factor_variable_1, y = continuous_variable_1)) + geom_boxplot() + labs(title = “Box Plot of Continuous Variable 1 by Factor Variable 1”, x = “Factor Variable 1”, y = “Continuous Variable 1”)

Documentation of Findings

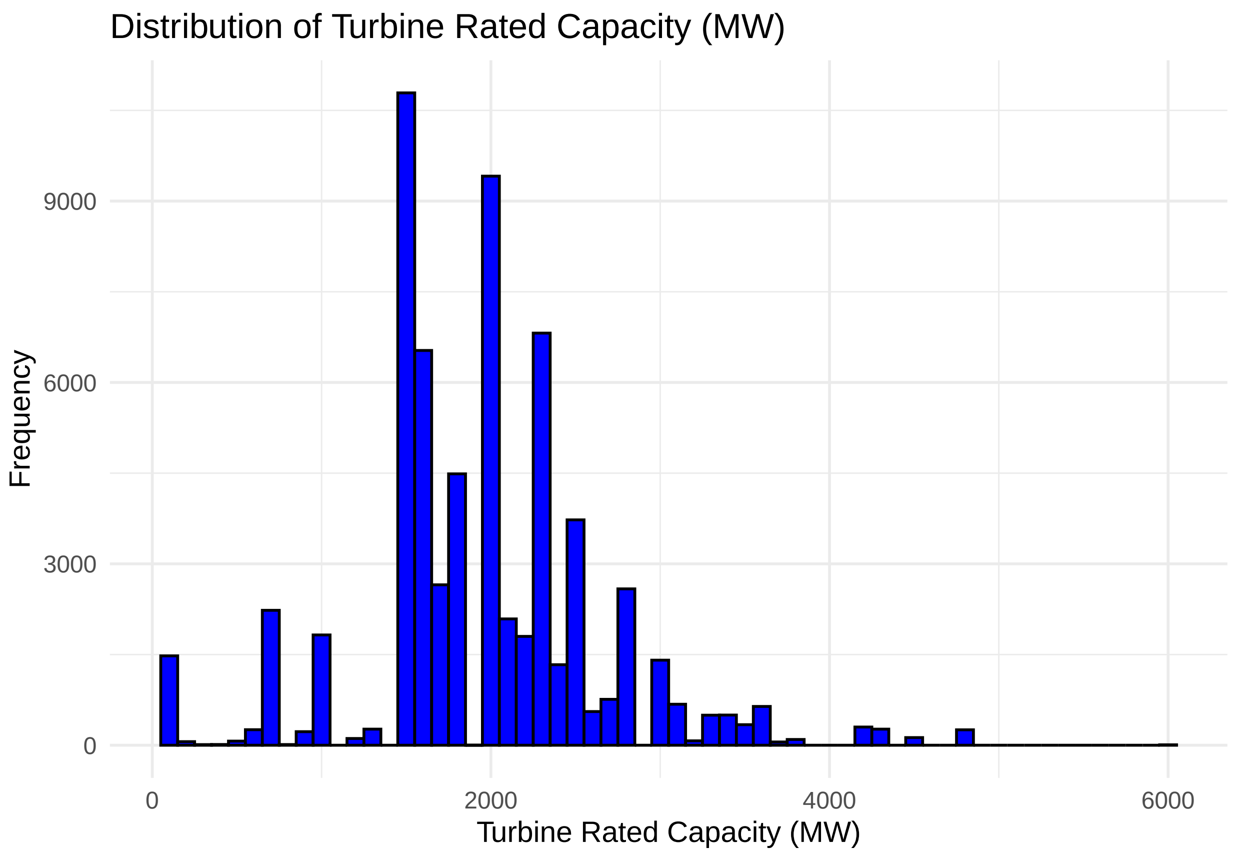
```
‘‘‘r
# Histograms for continuous variables
continuous_vars <- c("p_cap", "t_cap", "t_hh", "t_rd", "t_rsa", "t_ttlh")
continuous_titles <- c("Turbine Power Capacity (MW)", "Turbine Rated Capacity (MW)",
                       "Turbine Hub Height (m)", "Turbine Rotor Diameter (m)",
                       "Turbine Rotor Sweep Area (sq m)", "Total Turbine Height (m)")

binwidths <- c(10, 100, 5, 2.5, 500, 5)


for (i in 1:length(continuous_vars)) {
  var <- continuous_vars[i]
  title <- continuous_titles[i]
  binwidth <- binwidths[i]

  print(ggplot(data, aes_string(x = var)) +
          geom_histogram(binwidth = binwidth, fill = "blue", color = "black") +
          labs(x = title, y = "Frequency", title = paste("Distribution of", title)) +
          theme_minimal())
}
```

```
## Warning: ‘aes_string()‘ was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with ‘aes()‘.
## i See also ‘vignette("ggplot2-in-packages")‘ for more information.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

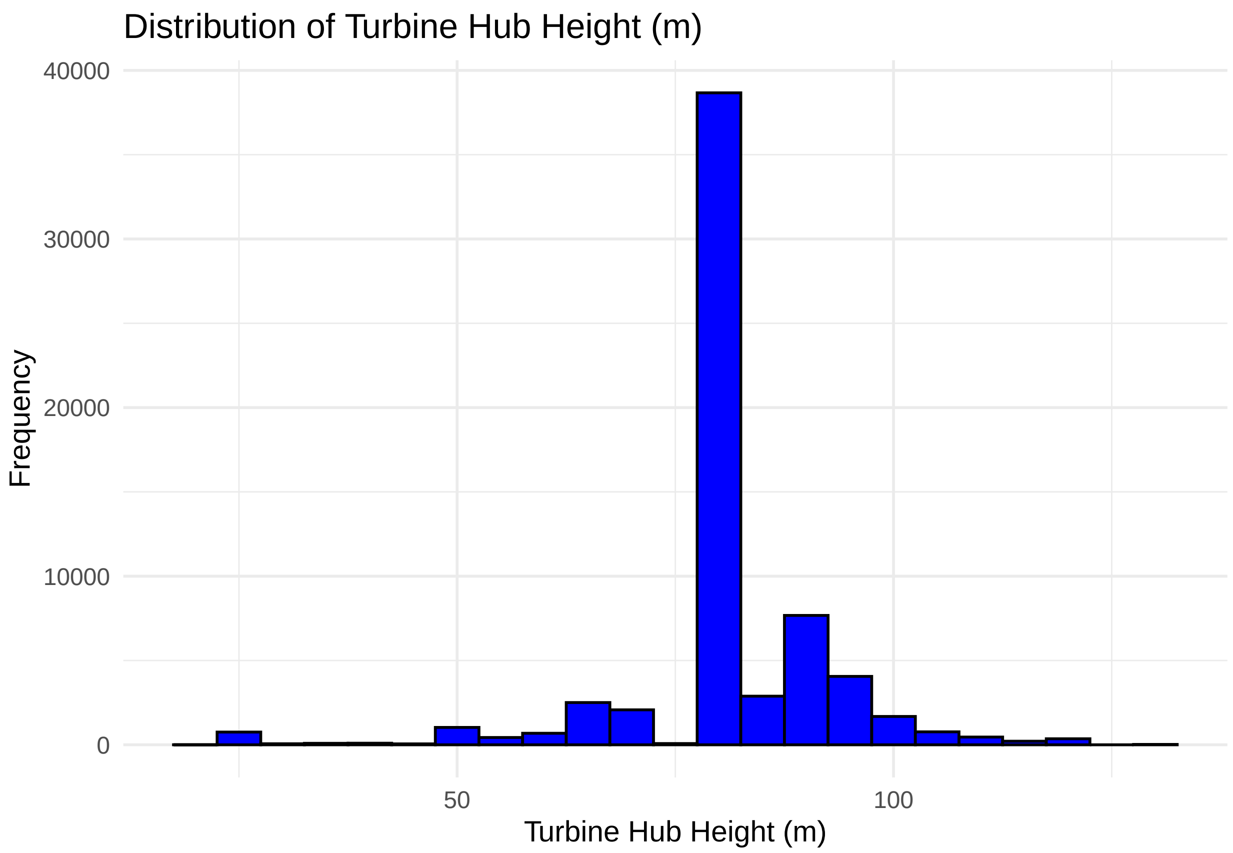
```
## Warning: Removed 4482 rows containing non-finite values (‘stat_bin()‘).
```



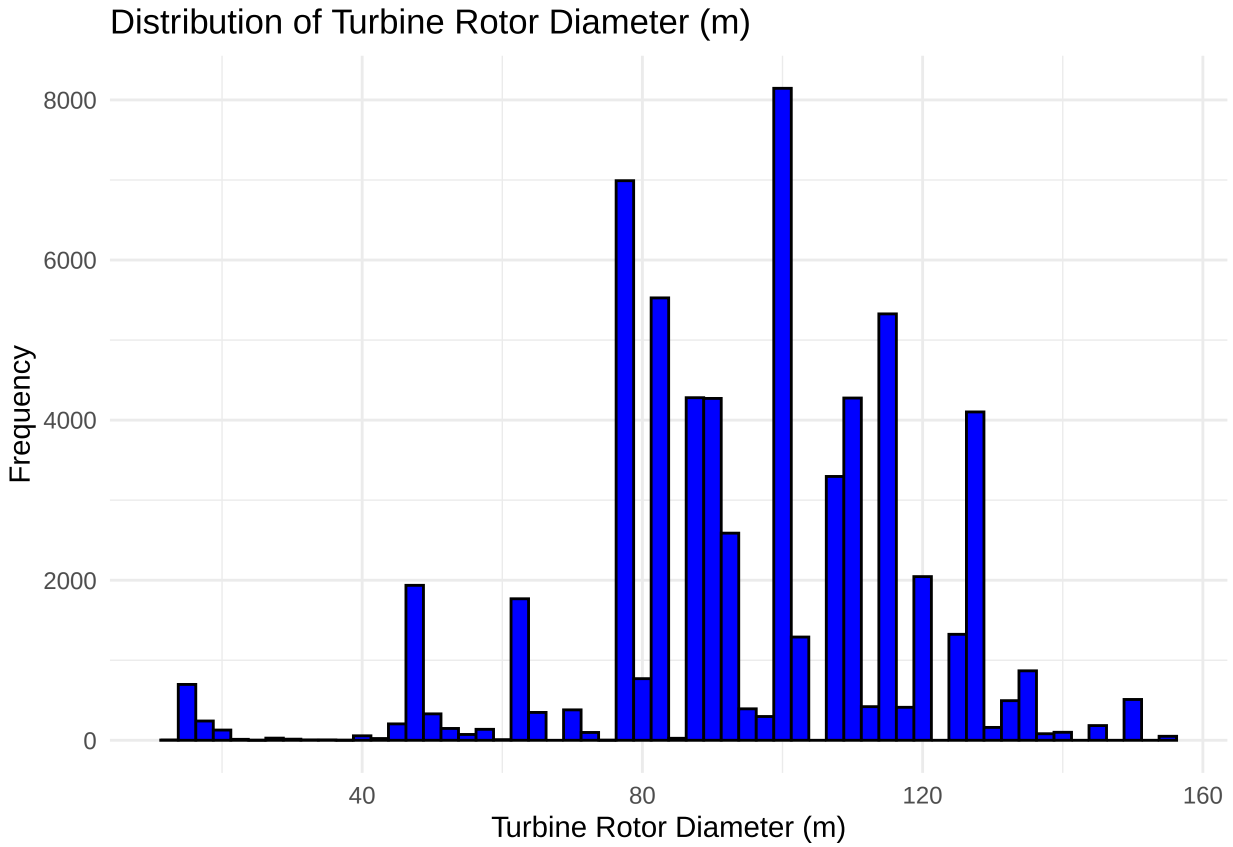
```
## Warning: Removed 5480 rows containing non-finite values (‘stat_bin()‘).
```



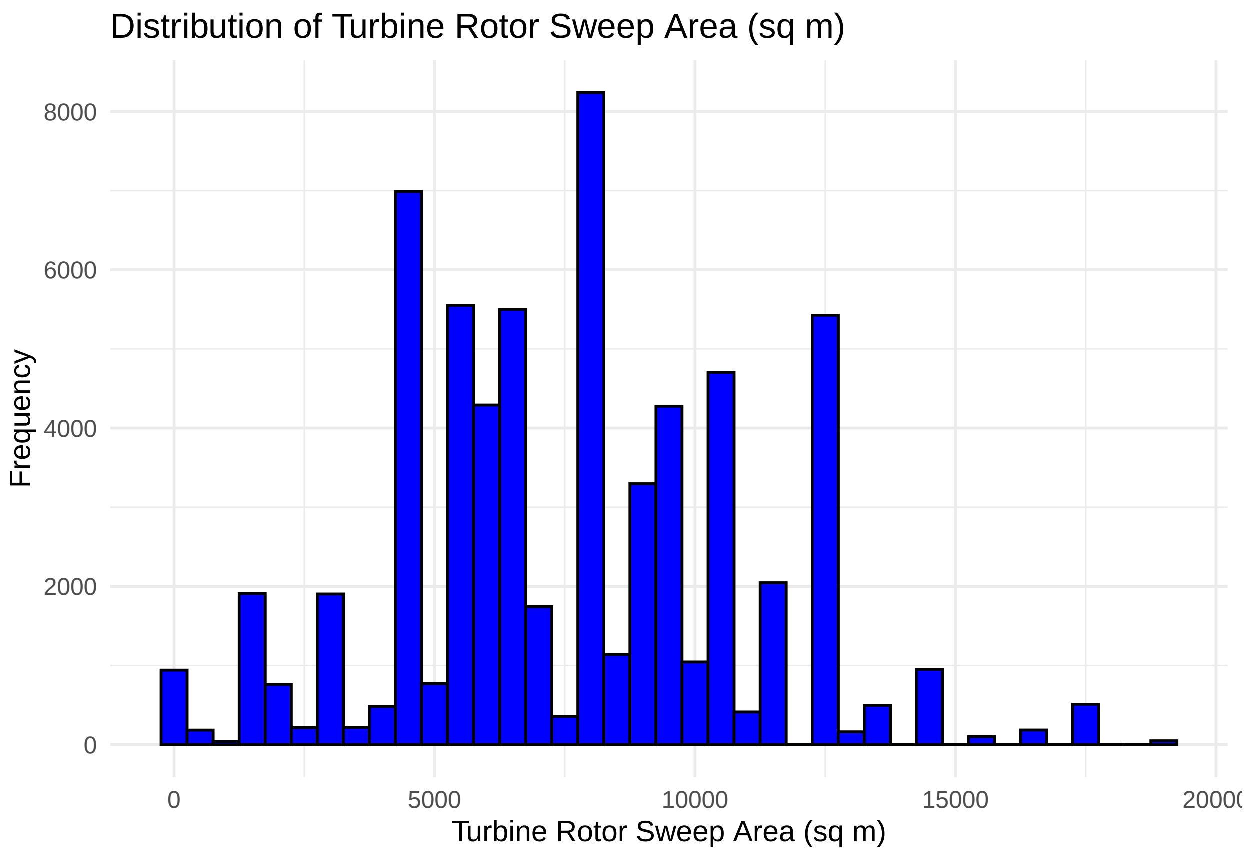
```
## Warning: Removed 6180 rows containing non-finite values (`stat_bin()`).
```



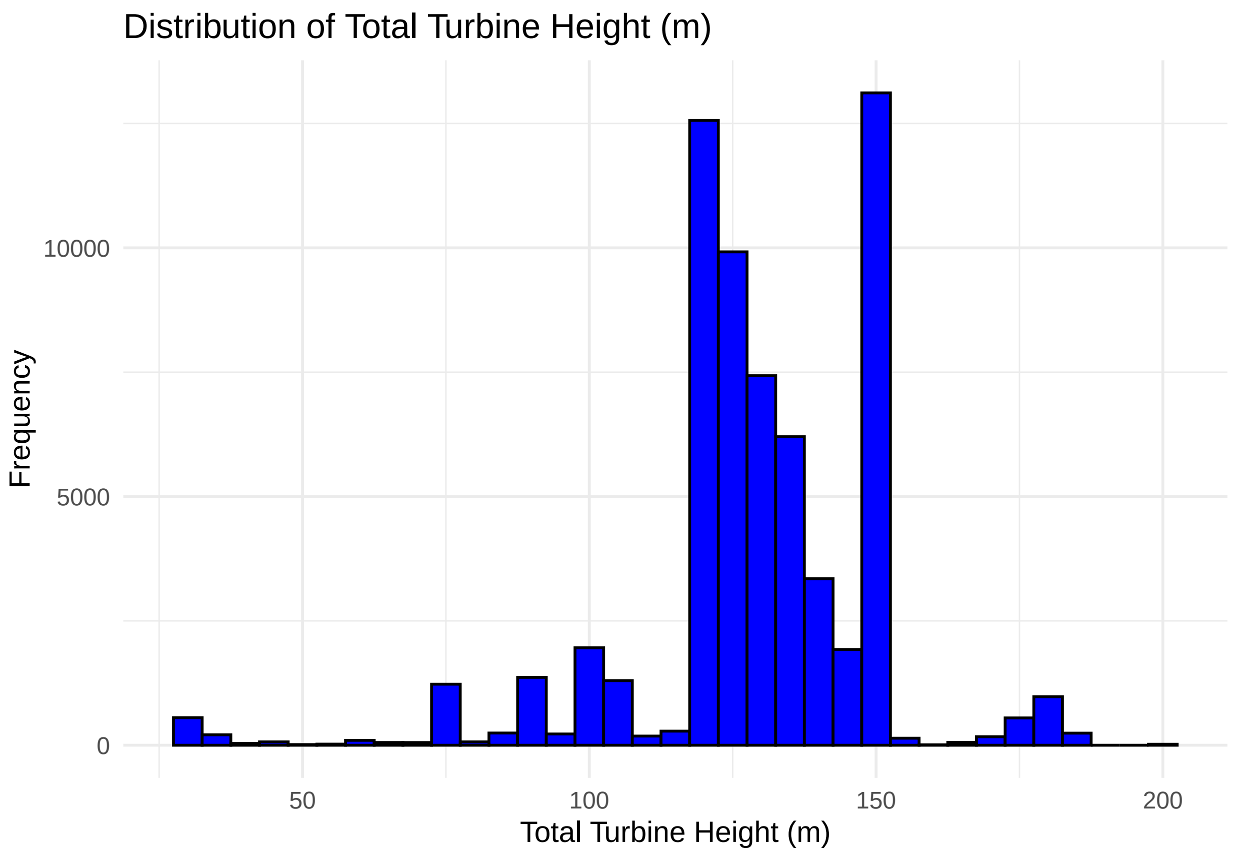
```
## Warning: Removed 5934 rows containing non-finite values (‘stat_bin()‘).
```



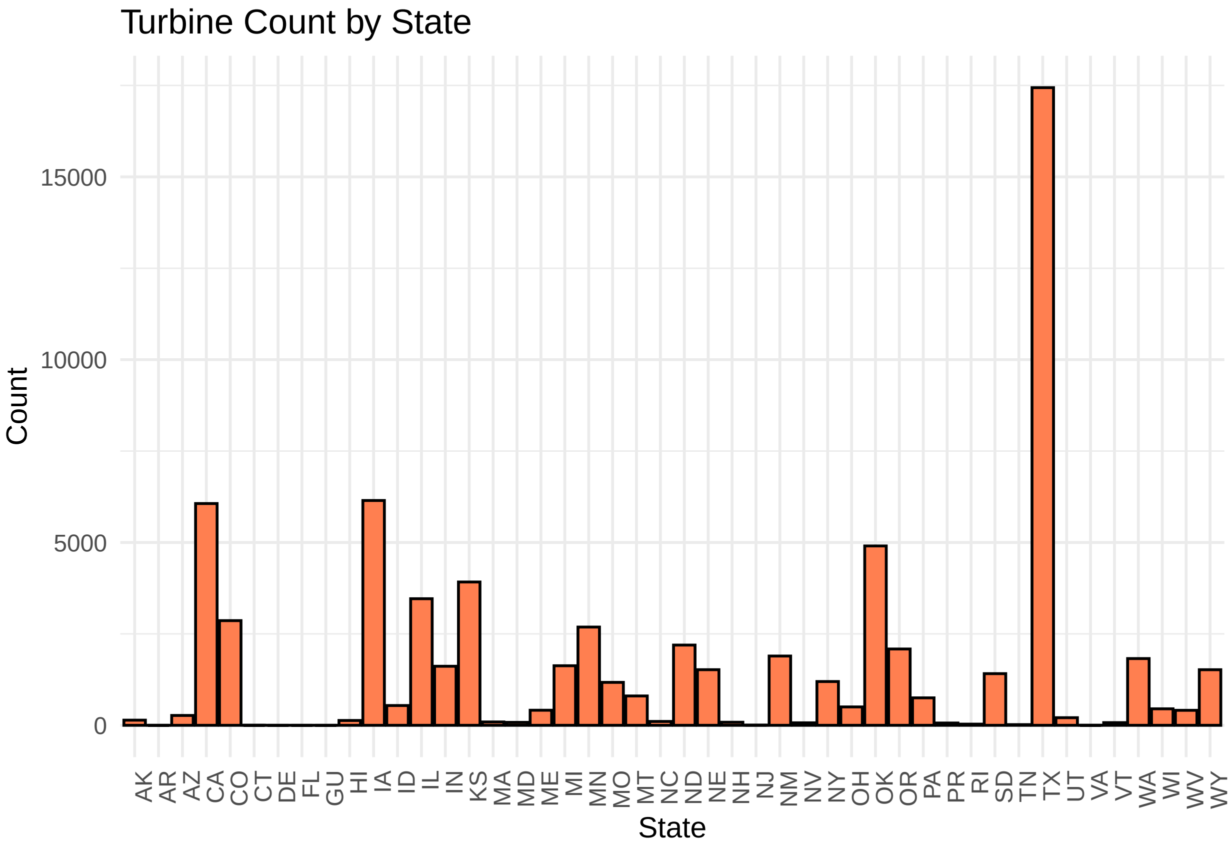
```
## Warning: Removed 5934 rows containing non-finite values (`stat_bin()`).
```

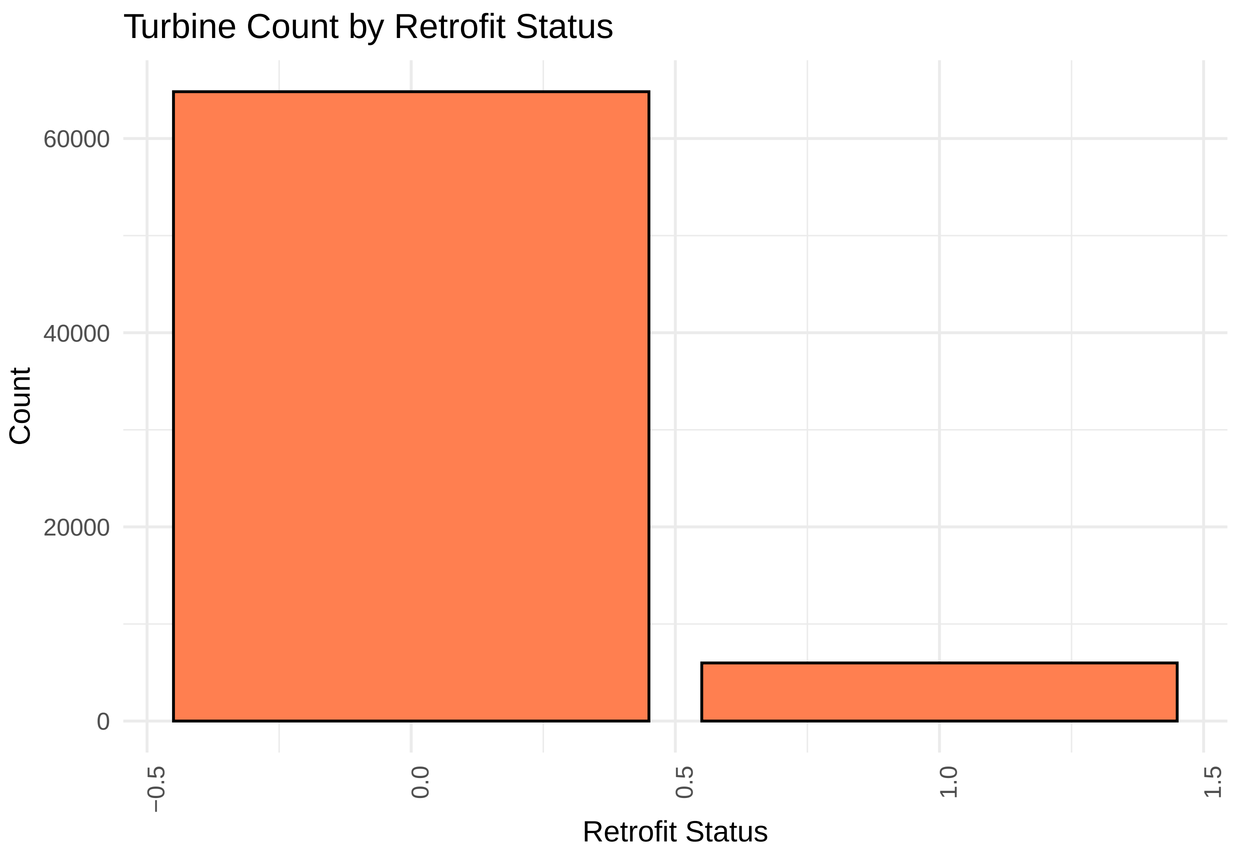


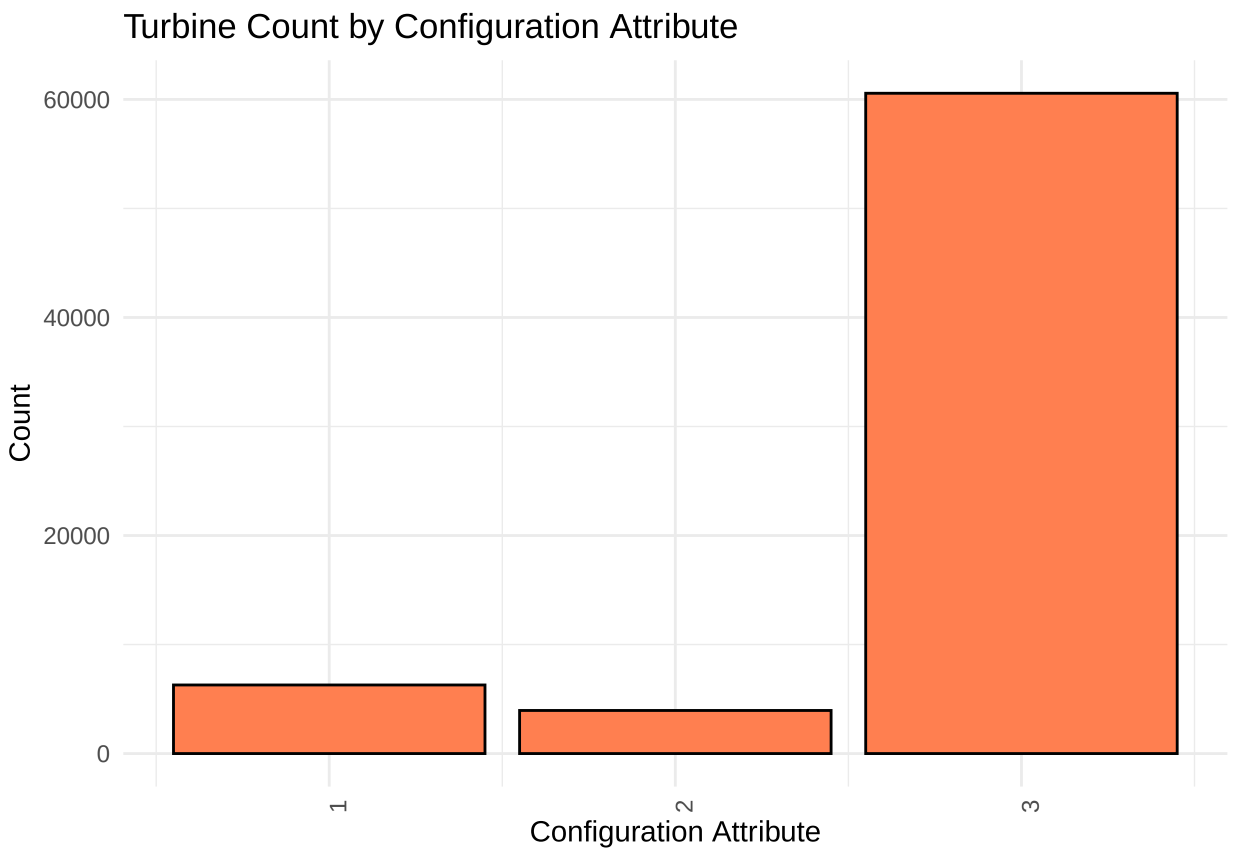
```
## Warning: Removed 6180 rows containing non-finite values (‘stat_bin()‘).
```

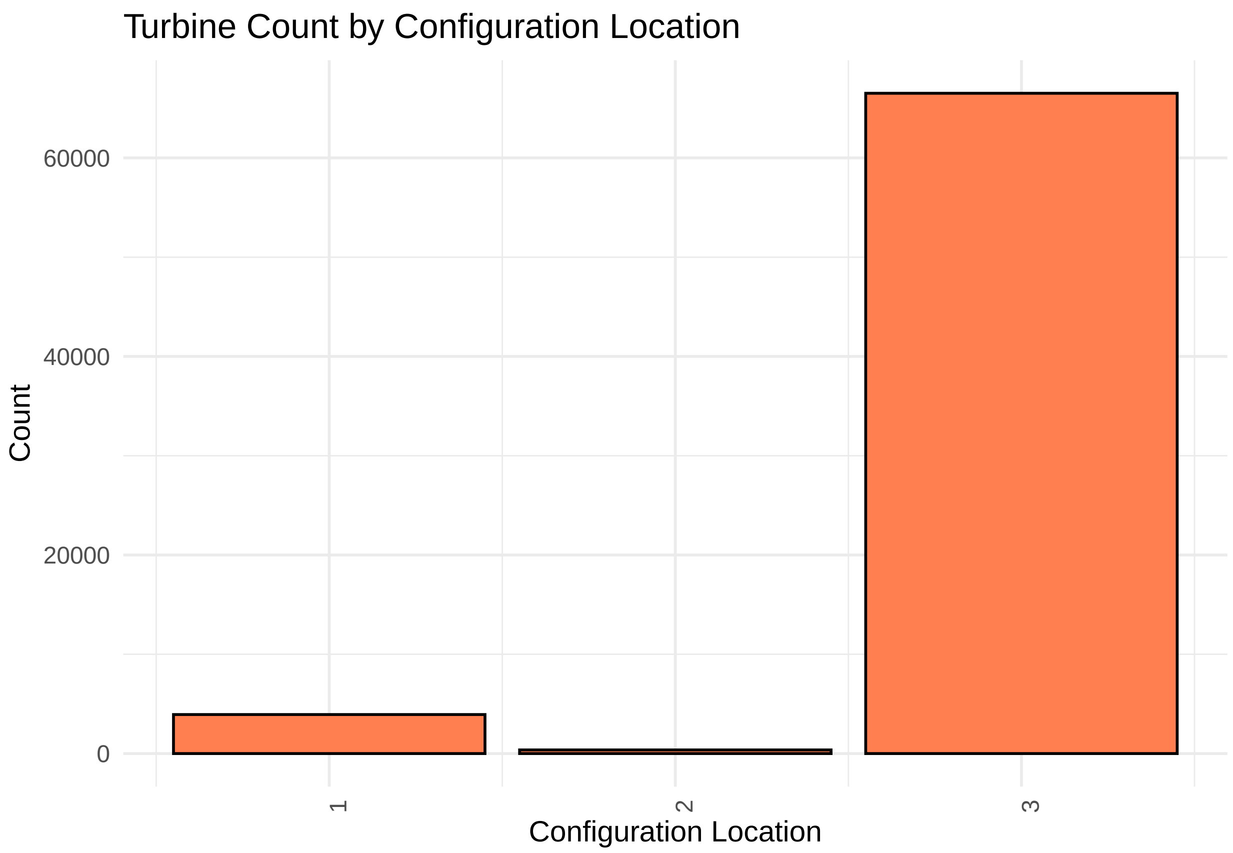


```
# Bar plots for categorical variables
categorical_vars <- c("t_state", "retrofit", "t_conf_atr", "t_conf_loc")
categorical_titles <- c("State", "Retrofit Status", "Configuration Attribute", "Configuration Location"

for (i in 1:length(categorical_vars)) {
  var <- categorical_vars[i]
  title <- categorical_titles[i]

  print(ggplot(data, aes_string(x = var)) +
          geom_bar(fill = "coral", color = "black") +
          labs(x = title, y = "Count", title = paste("Turbine Count by", title)) +
          theme_minimal() +
          theme(axis.text.x = element_text(angle = 90, hjust = 1)))
}
```

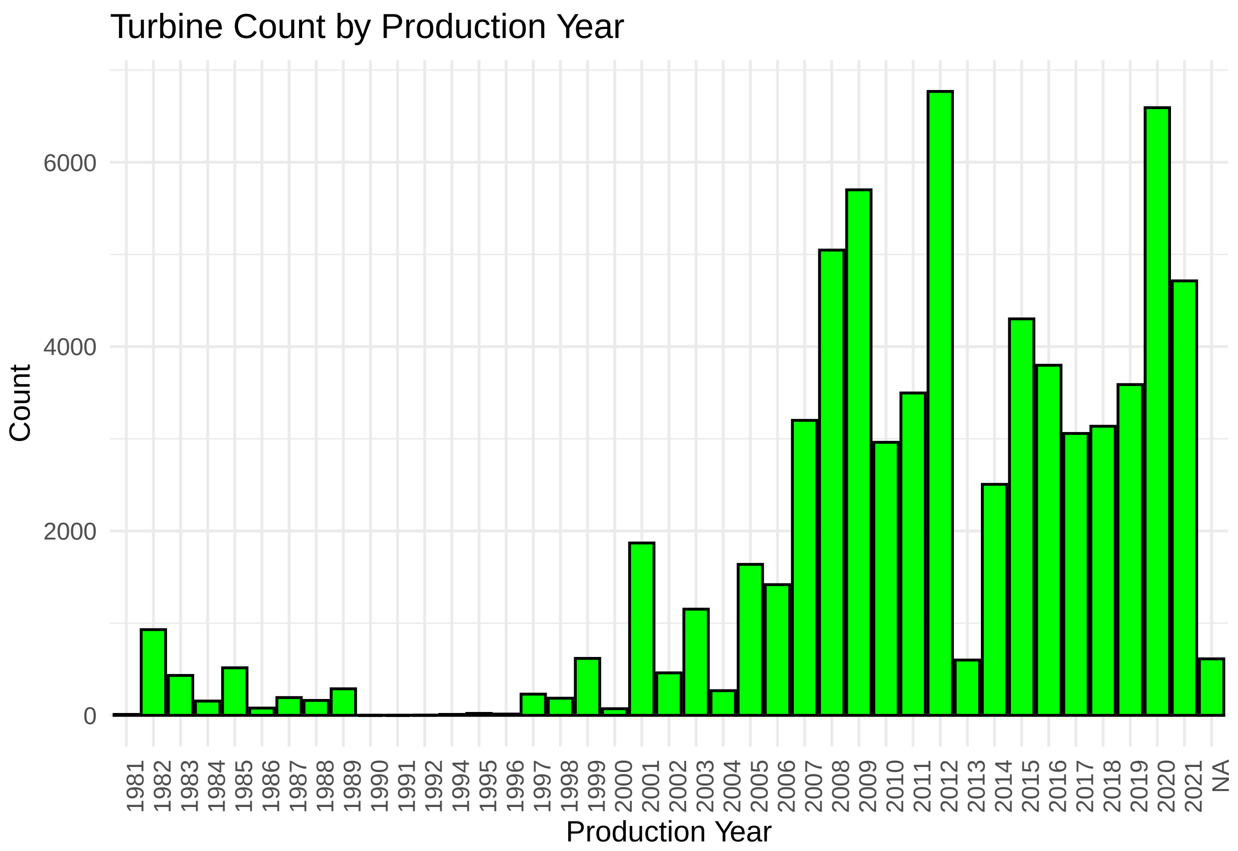

Turbine Count by State
Count
15000
10000
5000
0
AK
AR
AZ
CA
CO
CT
DE
FL
GU
HI
IA
ID
IL
IN
KS
MA
MD
ME
MI
MN
MO
MT
NC
ND
NE
NH
NJ
NM
NV
NY
OH
OK
OR
PA
PR
RI
SD
TN
TX
UT
VA
VT
WA
WI
WV
WY
State

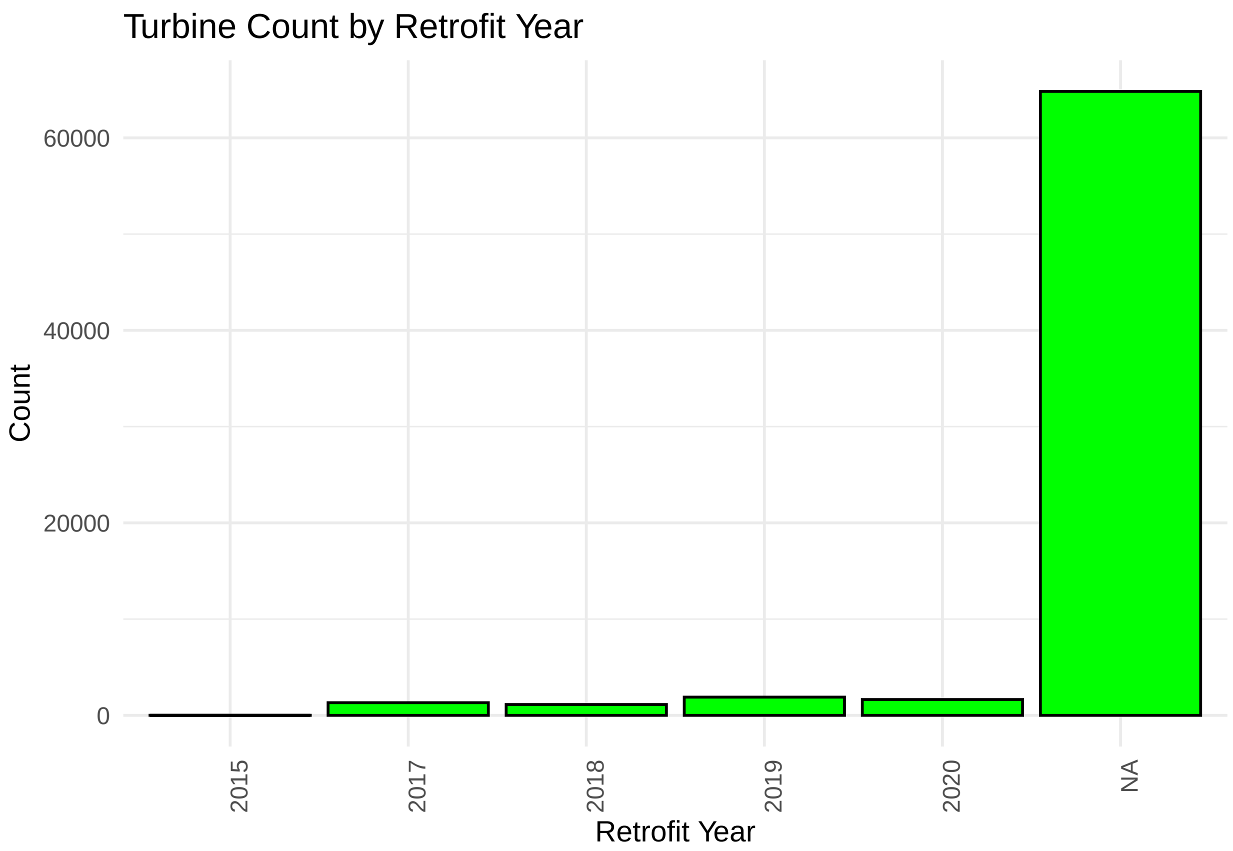

Turbine Count by Retrofit Status
Count
60000
40000
20000
0
−0.5
0.0
0.5
1.0
1.5
Retrofit Status


Turbine Count by Configuration Attribute
60000
40000
20000
0
Count
1
2
3
Configuration Attribute



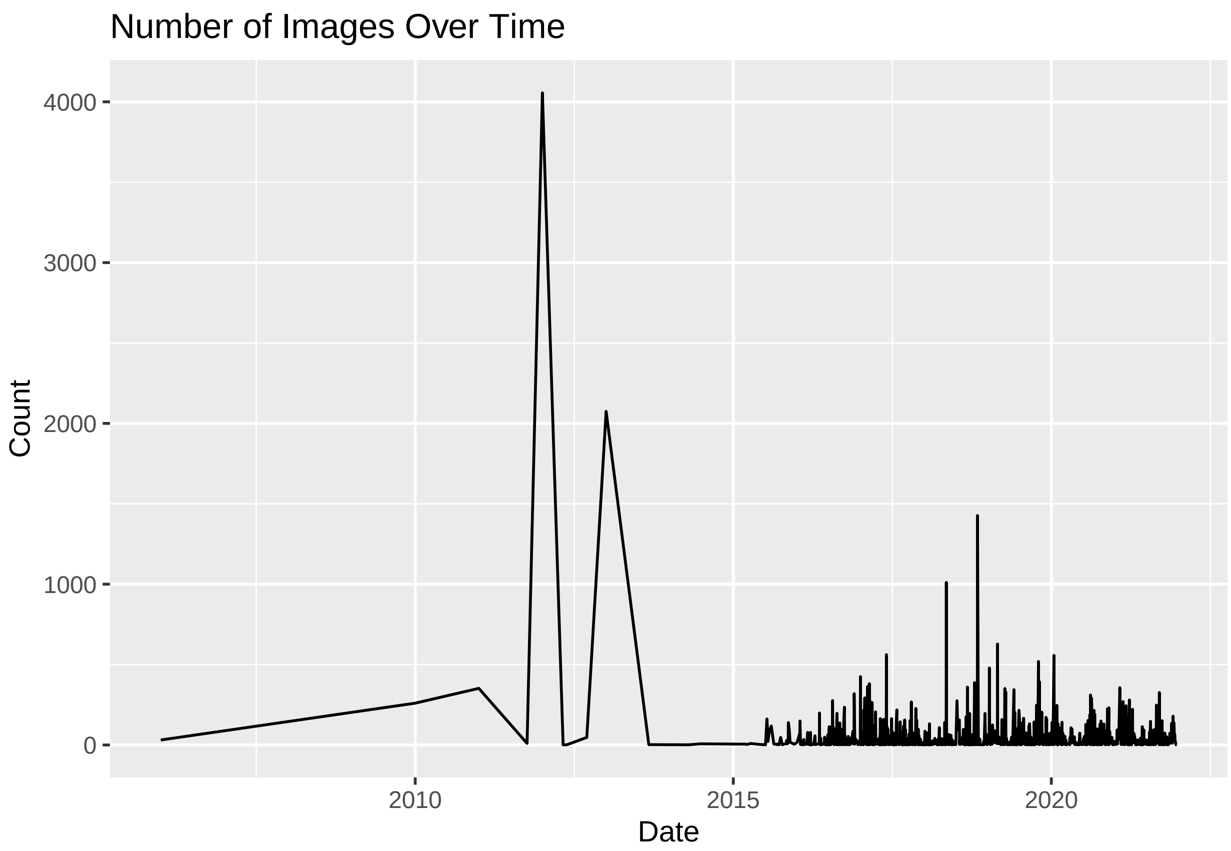
```
# Bar plot for year variables
year_vars <- c("p_year", "retrofit_year")
year_titles <- c("Production Year", "Retrofit Year")

for (i in 1:length(year_vars)) {
  var <- year_vars[i]
  title <- year_titles[i]

  data[[var]] <- as.factor(data[[var]])

  print(ggplot(data, aes_string(x = var)) +
          geom_bar(fill = "green", color = "black") +
          labs(x = title, y = "Count", title = paste("Turbine Count by", title)) +
          theme_minimal() +
          theme(axis.text.x = element_text(angle = 90, hjust = 1)))
}
```


Turbine Count by Production Year
Count
6000
4000
2000
0
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
NA
Production Year



```
# Convert date to Date format
data$t_img_date <- as.Date(data$t_img_date, format = "%m/%d/%Y")

# Line chart for date variable 't_img_date'
ggplot(data, aes(x = t_img_date)) +
  geom_line(stat = "count", aes(group = 1)) +
  labs(x = "Date", y = "Count", title = "Number of Images Over Time")
```

```
## Warning: Removed 8316 rows containing non-finite values (‘stat_count()‘).
```



```
register_stadiamaps("f94c64ea-35d9-425f-af7a-e139e3bd6242", write = TRUE)
```

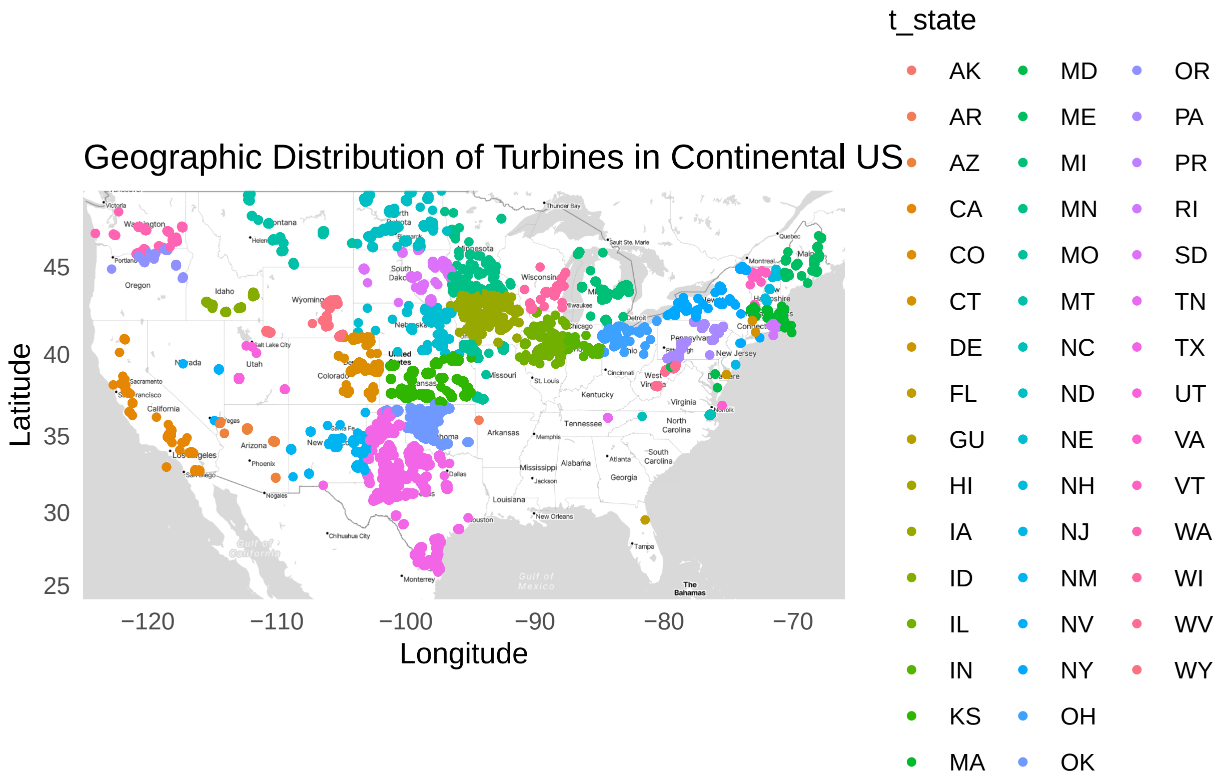
```
## i Replacing old key (f94c64ea) with new key in /Users/kyler/.Renviron
```

```
continental_bbox <- c(left = -125, bottom = 24, right = -66, top = 49)
continental_map <- get_stadiamap(bbox = continental_bbox, zoom = 5, maptype = "stamen_toner_lite")
```

```
## i © Stadia Maps © Stamen Design © OpenMapTiles © OpenStreetMap contributors.
```

```
ggmap(continental_map) +
  geom_point(data = data, aes(x = xlong, y = ylat, color = t_state), size = 1) +
  labs(x = "Longitude", y = "Latitude", title = "Geographic Distribution of Turbines in Continental US"
  theme_minimal()
```

```
## Warning: Removed 337 rows containing missing values (‘geom_point()‘).
```

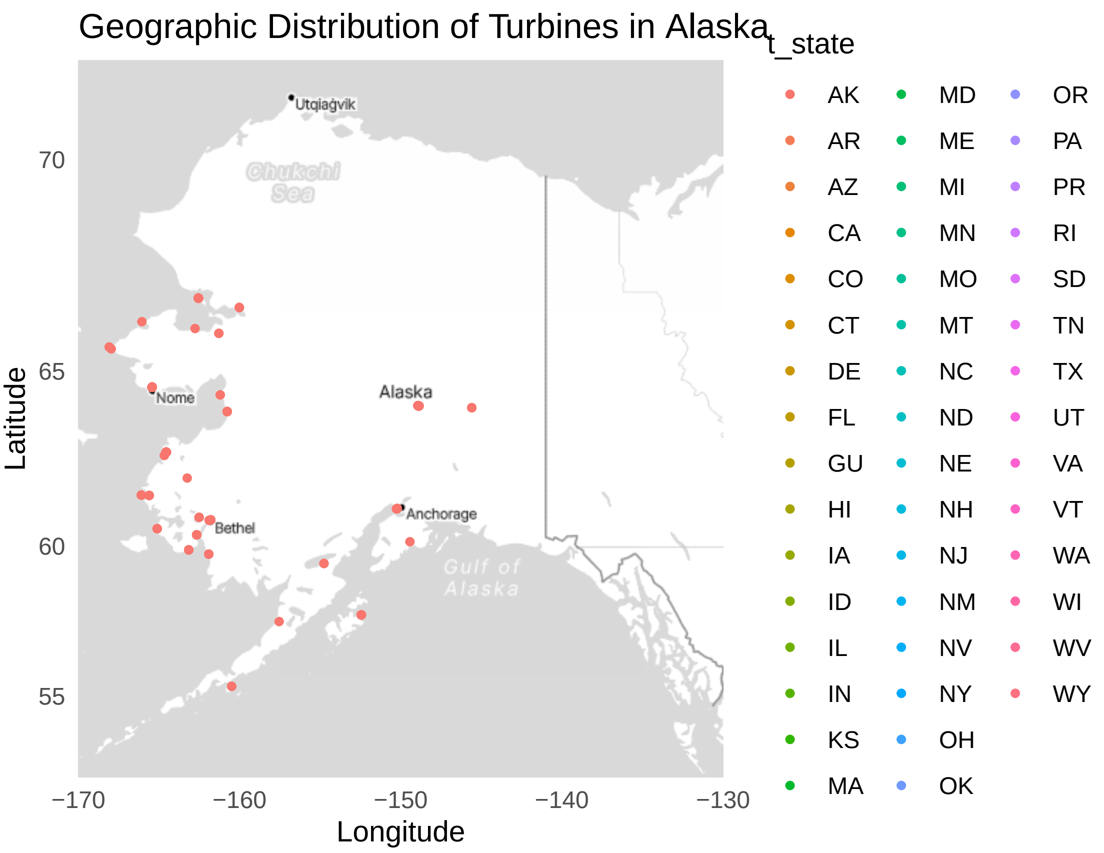


```
alaska_bbox <- c(left = -170, bottom = 52, right = -130, top = 72)
alaska_map <- get_stadiamap(bbox = alaska_bbox, zoom = 4, maptype = "stamen_toner_lite")
```

```
## i © Stadia Maps © Stamen Design © OpenMapTiles © OpenStreetMap contributors.
```

```
ggmap(alaska_map) +
  geom_point(data = data, aes(x = xlong, y = ylat, color = t_state), size = 1) +
  labs(x = "Longitude", y = "Latitude", title = "Geographic Distribution of Turbines in Alaska") +
  theme_minimal()
```

```
## Warning: Removed 70675 rows containing missing values (‘geom_point()‘).
```

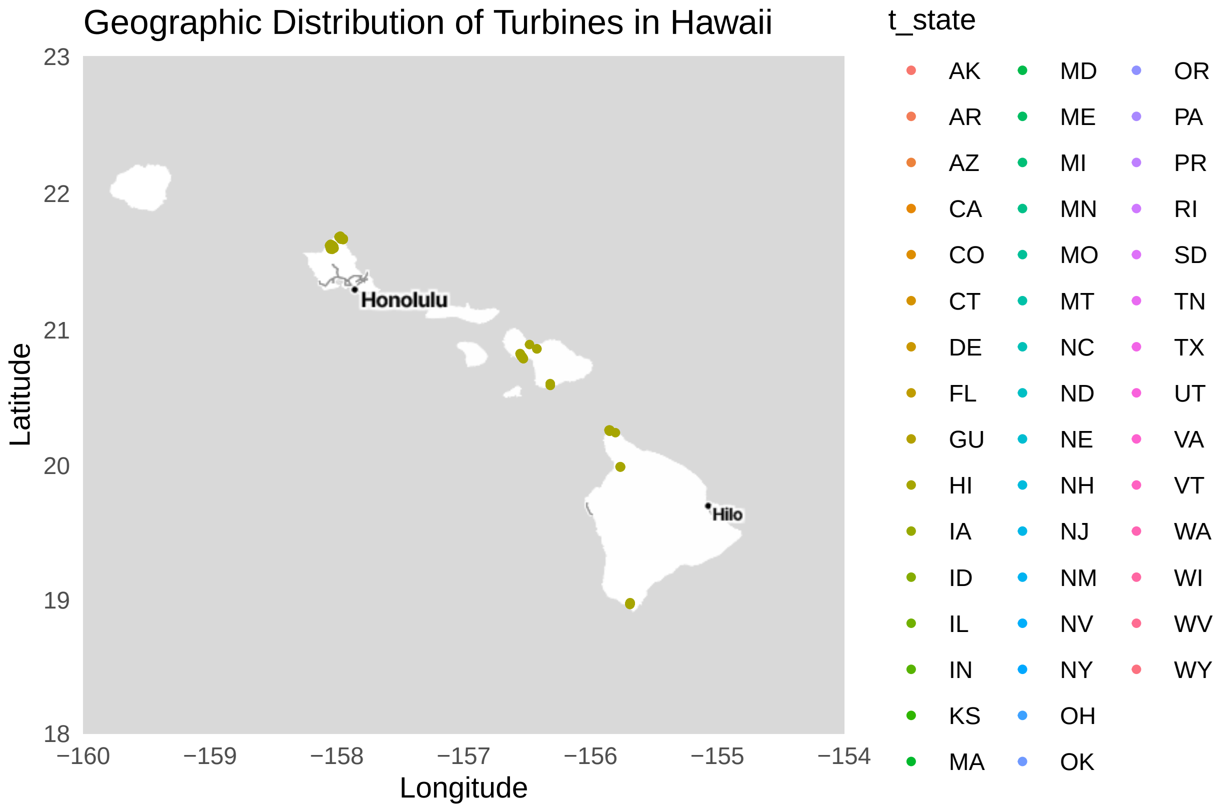


```r
hawaii_bbox <- c(left = -160, bottom = 18, right = -154, top = 23)
hawaii_map <- get_stadiamap(bbox = hawaii_bbox, zoom = 7, maptype = "stamen_toner_lite")
```

```
## i © Stadia Maps © Stamen Design © OpenMapTiles © OpenStreetMap contributors.
```

```r
ggmap(hawaii_map) +
  geom_point(data = data, aes(x = xlong, y = ylat, color = t_state), size = 1) +
  labs(x = "Longitude", y = "Latitude", title = "Geographic Distribution of Turbines in Hawaii") +
  theme_minimal()
```

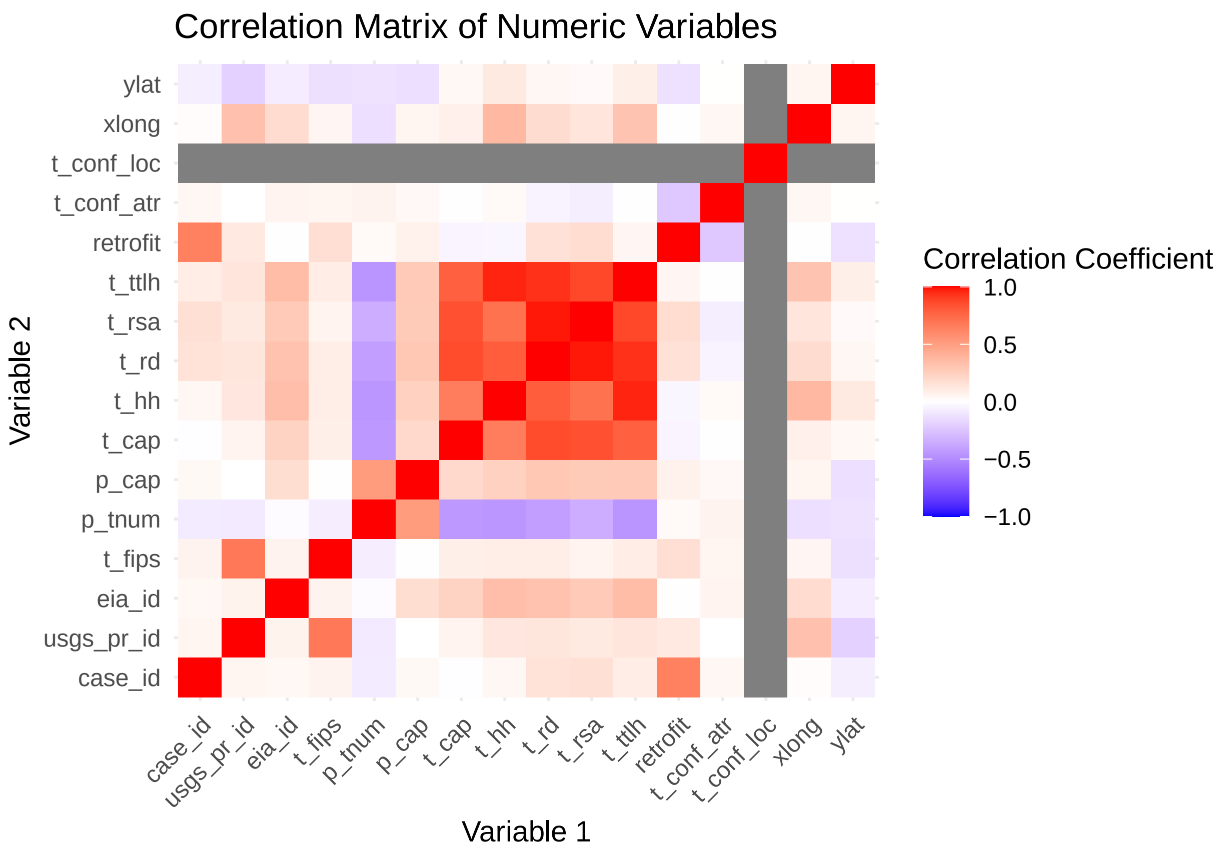
```
## Warning: Removed 70676 rows containing missing values (‘geom_point()‘).
```



```r
# Correlation plot if there are multiple numeric variables
numeric_data <- data %>% select_if(is.numeric)
correlation_matrix <- cor(numeric_data, use = "complete.obs")
```

```
## Warning in cor(numeric_data, use = "complete.obs"): the standard deviation is
## zero
```

```r
print(ggplot(data = as.data.frame(as.table(correlation_matrix)),
       aes(x = Var1, y = Var2, fill = Freq)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0, limit = c(-1,1)) +
  labs(title = "Correlation Matrix of Numeric Variables",
       x = "Variable 1",
       y = "Variable 2",
       fill = "Correlation Coefficient") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)))
```



```
# End of EDA section with session information
sessionInfo()
```

```
## R version 4.1.2 (2021-11-01)
## Platform: aarch64-apple-darwin20 (64-bit)
## Running under: macOS 14.2.1
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.1-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.1-arm64/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] ggmap_4.0.0     lubridate_1.9.3 forcats_1.0.0   stringr_1.5.1
##  [5] dplyr_1.1.4     purrr_1.0.2     readr_2.1.5     tidyr_1.3.1
##  [9] tibble_3.2.1    ggplot2_3.4.4   tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.12       highr_0.9         plyr_1.8.9        pillar_1.9.0
##  [5] compiler_4.1.2    bitops_1.0-7      tools_4.1.2       digest_0.6.29
##  [9] timechange_0.3.0  evaluate_0.14     lifecycle_1.0.4   gtable_0.3.0
## [13] png_0.1-8         pkgconfig_2.0.3   rlang_1.1.3       cli_3.6.2
## [17] rstudioapi_0.15.0 curl_5.2.0        yaml_2.2.1        xfun_0.29
## [21] fastmap_1.1.0     httr_1.4.7        withr_3.0.0       knitr_1.37
## [25] maps_3.4.2        generics_0.1.3    vctrs_0.6.5       hms_1.1.3
## [29] grid_4.1.2        tidyselect_1.2.0  glue_1.7.0        R6_2.5.1
## [33] jpeg_0.1-10       fansi_1.0.2       rmarkdown_2.11.12 farver_2.1.0
## [37] tzdb_0.4.0        magrittr_2.0.3    scales_1.3.0      htmltools_0.5.2
## [41] colorspace_2.0-3  labeling_0.4.2    utf8_1.2.2        stringi_1.7.6
## [45] munsell_0.5.0     crayon_1.4.2
```