Regression_weather

2023-02-09

import datasets

```r
df <- read.csv("merged_data.csv")
df <- na.omit(df)
```

```r
#add a route variable that indicates the start.station-end.station pair of the ride
locs <- unique(df$Start.station.number)
df$routes <- with(df, paste0(Start.station.number,"-",End.station.number))
unique_routes <- unique(with(df,paste0(Start.station.number,"-",End.station.number)))

#transform the time variable from chr to POSIXct
df$Start.date <- strptime(df$Start.date, format = "%Y-%m-%d %H:%M:%S")
df$Start.date <- as.POSIXct(df$Start.date, format = "%Y-%m-%d %H:%M:%S")

#drop irrelevant information
drops <-c("End.date","Start.station.number","Start.station","End.station.number","End.station", "Bike.nu
df <- df[ , !(names(df) %in% drops)]

#we set the first date in the dataset as the starting point
#measure how much time has elapsed since the first date (in seconds)
df$Start.date <- as.numeric(df$Start.date- df$Start.date[1])
```

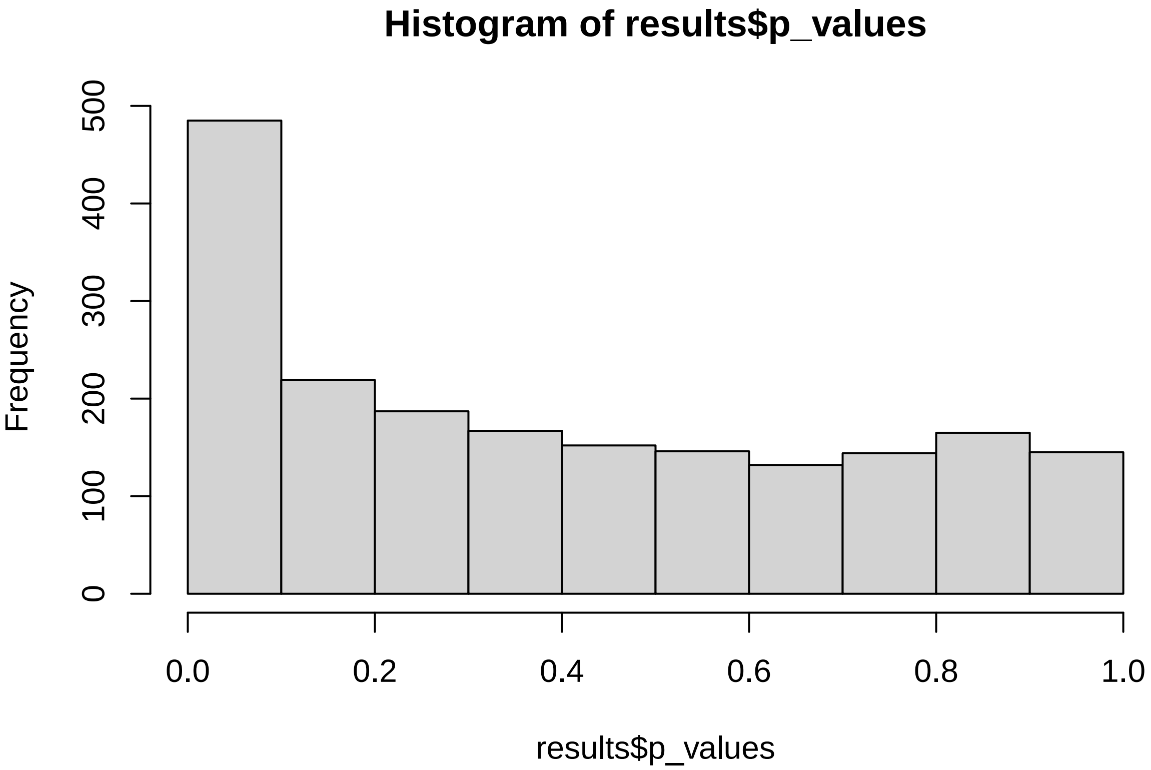
```r
#count the number of counts for each route and arrange from most to least
count_routes <- df %>%
  group_by(routes) %>%
  summarise(count = n()) %>%
  filter(count >= 200)
```

```r
regression_one_route <- function(route){
  df <- df %>%
    filter(routes == route)

  model <- lm(Duration ~ Start.date + Member.type + weekday + weather_description, data=df)
  pvals <- summary(model)$coefficients[2,4]
  return(pvals)
}
```

```r
regression_all_routes <- function(){
  size = length(count_routes$routes)
  p_values = matrix(0, size)
  for (i in 1:size){
    single_p_value <- regression_one_route(count_routes$routes[i])
    p_values[i] = single_p_value
  }
  count_routes$p_values = p_values
  return(count_routes)
}
```

```
results <- regression_all_routes()
hist(results$p_values)
```



```
write.csv(results, "p_values_weather_regression.csv", row.names=FALSE)
```