EV Power - Lab 4 Project Report

Example Solution 1

Part 0: libraries

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.1     v stringr   1.5.2
v ggplot2   4.0.0     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- 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 beco
```

Part 1: Defining Research Question

Chosen Question:Do states with higher renewable usage have lower average electricity prices?

Part 2: Data Preparation and Cleaning

```
renew_2021_raw <- read_csv("data/renew-use-2021.csv")
```

```
Rows: 260 Columns: 3
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (3): State, Energy_Source, Renewable_Use_2021

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
renew_2022_raw <- read_csv("data/renew-use-2022.csv")
```

```
Rows: 260 Columns: 3
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (3): State, Energy_Source, Renewable_Use_2022

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
renew_2023_raw <- read_csv("data/renew-use-2023.csv")
```

```
Rows: 260 Columns: 3
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (3): State, Energy_Source, Renewable_Use_2023

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
avg_price_2021_to_2023_raw <- read_csv("data/av-energy-price-2021-2023.csv")
```

```
Rows: 54 Columns: 1
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (1): Total energy average price, dollars per million Btu,,,

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
renew_2021_clean <- renew_2021_raw |>
    mutate(
        across(Renewable_Use_2021, ~ str_replace_all(.,"[A-Z|a-z|~|\\$|\\.| ]", "")),
        Renewable_Use_2021 = as.numeric(Renewable_Use_2021)
        )|>
    group_by(State)|>
    summarise(total_use_2021 = sum(Renewable_Use_2021))|>
    arrange(desc(total_use_2021))|>
    slice(-1)
renew_2021_clean
```

```
# A tibble: 51 x 2
   State total_use_2021
   <chr>          <dbl>
 1 CA            810020
 2 TX            654199
 3 WA            394052
 4 IA            389786
 5 FL            297290
 6 GA            289113
 7 NY            263978
 8 AL            239816
 9 OR            225543
10 IL            224107
# i 41 more rows
```

```r
#renewable 2022
renew_2022_clean <- renew_2022_raw |>
    mutate(
        across(Renewable_Use_2022, ~ str_replace_all(.,"[A-Z|a-z|~|\\$|\\.| ]", "")),
        Renewable_Use_2022 = as.numeric(Renewable_Use_2022)
        )|>
    group_by(State)|>
    summarise(total_use_2022 = sum(Renewable_Use_2022))|>
    arrange(desc(total_use_2022))|>
    slice(-1)
renew_2022_clean
```

```
# A tibble: 51 x 2
   State total_use_2022
   <chr>          <dbl>
```

```
 1 CA            880995
 2 TX            751680
 3 IA            421784
 4 WA            418470
 5 FL            304605
 6 GA            293238
 7 NY            269883
 8 IL            248541
 9 OR            237768
10 AL            232035
# i 41 more rows
```

```
#renewable 2023
renew_2023_clean <- renew_2023_raw |>
    mutate(
        across(Renewable_Use_2023, ~ str_replace_all(.,"[A-Z|a-z|~|\\$|\\.| ]", "")),
        Renewable_Use_2023 = as.numeric(Renewable_Use_2023)
        )|>
    group_by(State)|>
    summarise(total_use_2023 = sum(Renewable_Use_2023))|>
    arrange(desc(total_use_2023))|>
    slice(-1)
renew_2023_clean
```

```
# A tibble: 51 x 2
   State total_use_2023
   <chr>          <dbl>
 1 CA           1065179
 2 tx            791210
 3 IA            414801
 4 wa            365955
 5 GA            291462
 6 fl            286307
 7 ny            272968
 8 il            245703
 9 OR            236063
10 Mn            223864
# i 41 more rows
```

```
avg_price_2021_to_2023_raw
```

```
# A tibble: 54 x 1
```

```
   `Total energy average price, dollars per million Btu,,,`
   <chr>
 1 ,,,
 2 State,2021,2022,2023
 3 AK,$20.03 per MMBtu,$27.33,$23.84 est.
 4 AL,about 17.85 USD,23.37 USD, 21.11
 5 AR,$18.42,$23.84 per MMBtu,$21.76
 6 AZ, 25.07,31.72 USD,about 30.28
 7 CA,$28.44,$37.35,$35.72 per MMBtu
 8 CO,20.64 USD, 25.85,23.85
 9 CT,about $25.85,$33.15,$32.32 est.
10 DC, 25.67,$30.84,about 32.28 USD
# i 44 more rows
```

```
# Try to filter the renewable column to only digit
colnames(avg_price_2021_to_2023_raw) <- c('x')

header_row <- avg_price_2021_to_2023_raw[2,1]
column_names <- str_split(header_row, ',')[[1]]
col1 <- column_names[1]
col1
```

```
[1] "State"
```

```
avg_price_2021_to_2023_clean <- avg_price_2021_to_2023_raw |>
    slice(3:nrow(avg_price_2021_to_2023_raw))|>
    mutate(
        across(everything(), ~ str_replace_all(.,"\\$", "")),
        across(everything(), ~ str_replace_all(.,"about","")),
        across(everything(), ~ str_replace_all(.,"per MMBtu", "")),
        across(everything(), ~str_replace_all(.,"USD","")),
        across(everything(), ~str_replace_all(.," ", "")),
        across(everything(), ~str_replace_all(.,"est\\.","")),
        across(everything(), ~str_replace_all(.," ",""))
        )|>
    mutate(
    col1 = str_extract(x, "^[A-Z]{2}"),
    col2 = str_extract(x, "(?<=^[A-Z]{2},)[^,]+"),
    col3 = str_extract(x, "(?<=,)[^,]+(?=,[^,]+$)"),
    col4 = str_extract(x, "[^,]+$"),
    col2 = as.numeric(col2),
    col3 = as.numeric(col3),
```

```
    col4 = as.numeric(col4)
  ) |>
  select(-1)
  colnames(avg_price_2021_to_2023_clean) <- column_names

avg_price_2021_to_2023_clean
```

```
# A tibble: 52 x 4
   State `2021` `2022` `2023`
   <chr>  <dbl>  <dbl>  <dbl>
 1 AK      20.0   27.3   23.8
 2 AL      17.8   23.4   21.1
 3 AR      18.4   23.8   21.8
 4 AZ      25.1   31.7   30.3
 5 CA      28.4   37.4   35.7
 6 CO      20.6   25.8   23.8
 7 CT      25.8   33.2   32.3
 8 DC      25.7   30.8   32.3
 9 DE      21.8   27.7   26.7
10 FL      22.5   29.4   28.1
# i 42 more rows
```

Part 3: Joining / Pivoting Datasets for Analysis

```
#For 2021 table
avg_price_2021 <- avg_price_2021_to_2023_clean|>
    select('State','2021')|>
    rename('Average price 2021' = '2021')
avg_price_2021
```

```
# A tibble: 52 x 2
   State `Average price 2021`
   <chr>                <dbl>
 1 AK                    20.0
 2 AL                    17.8
 3 AR                    18.4
 4 AZ                    25.1
 5 CA                    28.4
 6 CO                    20.6
```

```
 7 CT                    25.8
 8 DC                    25.7
 9 DE                    21.8
10 FL                    22.5
# i 42 more rows
```

```
tbl_2021 <- left_join(renew_2021_clean, avg_price_2021)
```

```
Joining with `by = join_by(State)`
```

```
tbl_2021
```

```
# A tibble: 51 x 3
   State total_use_2021 `Average price 2021`
   <chr>          <dbl>                <dbl>
 1 CA            810020                 28.4
 2 TX            654199                 16.4
 3 WA            394052                 21.0
 4 IA            389786                 16.4
 5 FL            297290                 22.5
 6 GA            289113                 19.8
 7 NY            263978                 22.6
 8 AL            239816                 17.8
 9 OR            225543                 21.5
10 IL            224107                 18.4
# i 41 more rows
```

```
#For 2022 table
avg_price_2022 <- avg_price_2021_to_2023_clean|>
    select('State','2022')|>
    rename('Average price 2022' = '2022')
avg_price_2022
```

```
# A tibble: 52 x 2
   State `Average price 2022`
   <chr>                <dbl>
 1 AK                    27.3
 2 AL                    23.4
 3 AR                    23.8
 4 AZ                    31.7
```

```
 5 CA                    37.4
 6 CO                    25.8
 7 CT                    33.2
 8 DC                    30.8
 9 DE                    27.7
10 FL                    29.4
# i 42 more rows
```

```
tbl_2022 <- left_join(renew_2022_clean, avg_price_2022)
```

```
Joining with `by = join_by(State)`
```

```
tbl_2022
```

```
# A tibble: 51 x 3
   State total_use_2022 `Average price 2022`
   <chr>          <dbl>                <dbl>
 1 CA            880995                 37.4
 2 TX            751680                 20.8
 3 IA            421784                 20.5
 4 WA            418470                 26.9
 5 FL            304605                 29.4
 6 GA            293238                 25.5
 7 NY            269883                 29.1
 8 IL            248541                 24.0
 9 OR            237768                 26.9
10 AL            232035                 23.4
# i 41 more rows
```

```
#For 2023 table
avg_price_2023 <- avg_price_2021_to_2023_clean|>
    select('State','2023')|>
    rename('Average price 2023' = '2023')
avg_price_2023
```

```
# A tibble: 52 x 2
   State `Average price 2023`
   <chr>                <dbl>
 1 AK                    23.8
 2 AL                    21.1
```

```
 3 AR                    21.8
 4 AZ                    30.3
 5 CA                    35.7
 6 CO                    23.8
 7 CT                    32.3
 8 DC                    32.3
 9 DE                    26.7
10 FL                    28.1
# i 42 more rows
```

```
tbl_2023 <- left_join(renew_2023_clean, avg_price_2023)
```

```
Joining with `by = join_by(State)`
```

```
tbl_2023
```

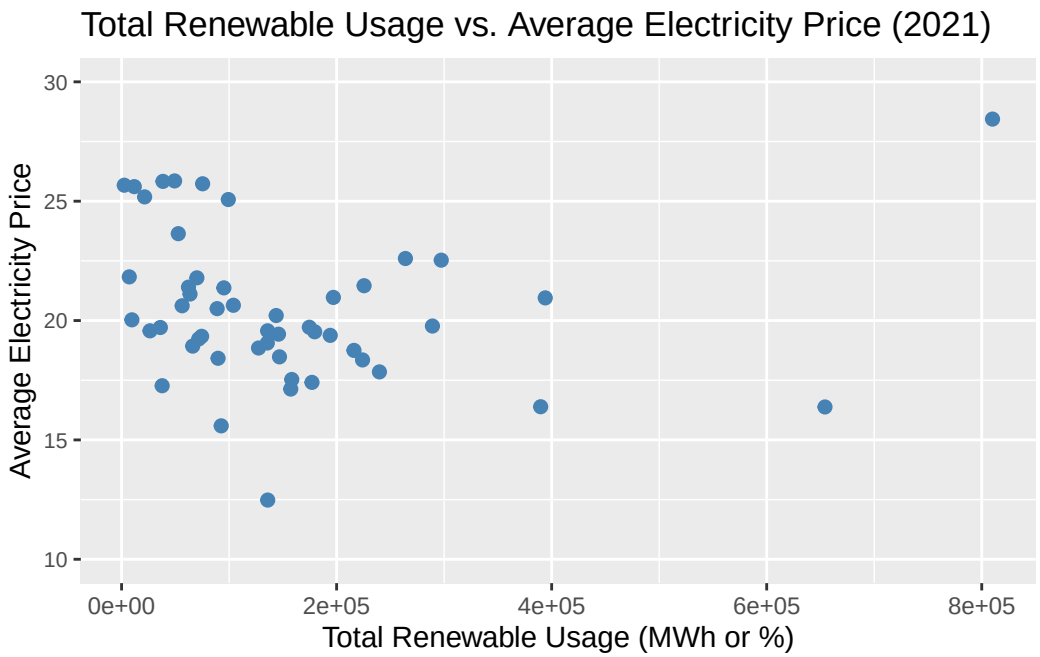
```
# A tibble: 51 x 3
   State total_use_2023 `Average price 2023`
   <chr>          <dbl>                <dbl>
 1 CA            1065179                 35.7
 2 tx             791210                 NA
 3 IA             414801                 18.1
 4 wa             365955                 NA
 5 GA             291462                 23.0
 6 fl             286307                 NA
 7 ny             272968                 NA
 8 il             245703                 NA
 9 OR             236063                 26.6
10 Mn             223864                 NA
# i 41 more rows
```

Part 4: Mapping Visualization

```
#Map 2021
ggplot(tbl_2021, aes(x = total_use_2021, y = `Average price 2021`)) +
  geom_point(color = "steelblue", size = 2) +
  labs(
    title = "Total Renewable Usage vs. Average Electricity Price (2021)",
    x = "Total Renewable Usage (MWh or %)",
```

```
    y = "Average Electricity Price"
  ) +
  scale_y_continuous(
    breaks = c(10, 15, 20, 25, 30),
    limits = c(10, 30)
  ) +
  theme(axis.text.y = element_text(size = 8))
```

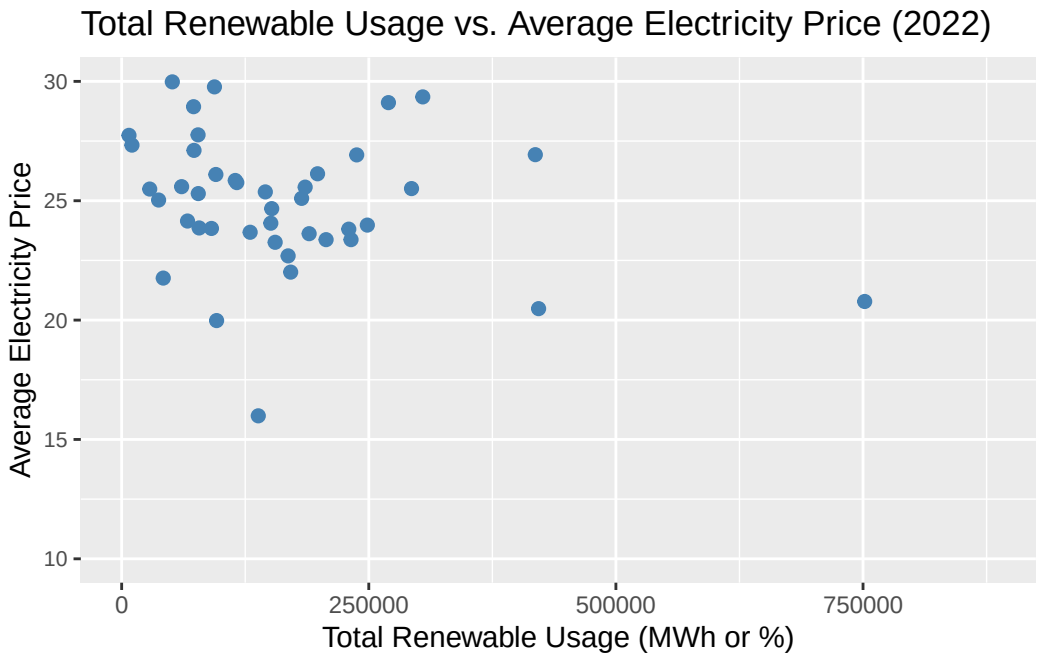
```
Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_point()`).
```




```
#Map 2022
ggplot(tbl_2022, aes(x = total_use_2022, y = `Average price 2022`)) +
  geom_point(color = "steelblue", size = 2) +
  labs(
    title = "Total Renewable Usage vs. Average Electricity Price (2022)",
    x = "Total Renewable Usage (MWh or %)",
    y = "Average Electricity Price"
  ) +
  scale_y_continuous(
    breaks = c(10, 15, 20, 25, 30),
```

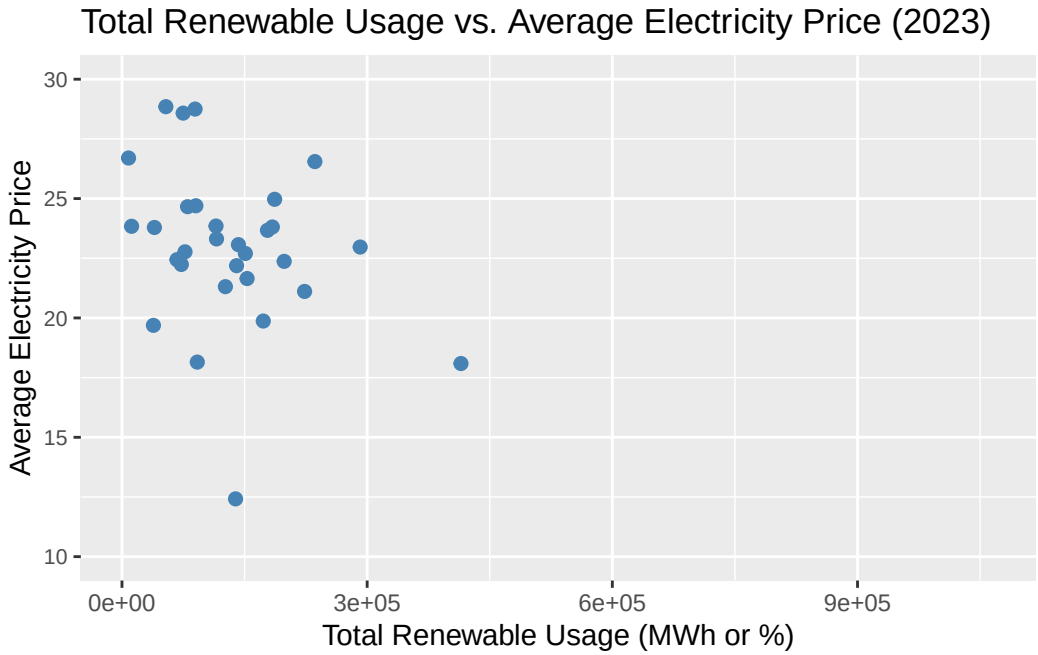
```
    limits = c(10, 30)
  ) +
  theme(axis.text.y = element_text(size = 8))
```

```
Warning: Removed 9 rows containing missing values or values outside the scale range
(`geom_point()`).
```




```
#Map 2023
ggplot(tbl_2023, aes(x = total_use_2023, y = `Average price 2023`)) +
  geom_point(color = "steelblue", size = 2) +
  labs(
    title = "Total Renewable Usage vs. Average Electricity Price (2023)",
    x = "Total Renewable Usage (MWh or %)",
    y = "Average Electricity Price"
  ) +
  scale_y_continuous(
    breaks = c(10, 15, 20, 25, 30),
    limits = c(10, 30)
  ) +
  theme(axis.text.y = element_text(size = 8))
```

```
Warning: Removed 21 rows containing missing values or values outside the scale range
(`geom_point()`).
```



Part 5: Final Deliverable

To answer this question, I worked on cleaning and joining two datasets: renewable energy usage and average electricity price for each state and year. Since the average price table was messy, I created a new data frame and separated the state names and years into different columns. After that, I joined the average electricity price table with the renewable usage table by state for each year.Then, I created a scatter plot to visualize whether states with higher renewable usage tend to have lower electricity prices. The plot shows that most states with lower renewable usage tend to have higher electricity prices. However, California, which has the highest renewable usage, does not have the lowest electricity price. This suggests that other factors, such as temperature, population density, or household size, may also affect electricity prices.