Climate resilience requires equitable access to quality green energy jobs. The City of Saint Paul is at the forefront.

Elham Ali

2024-09-19

Abstract

Minnesota, particularly the City of Saint Paul, has seen a surge in climate resilience funding aimed at expanding green energy job opportunities. However, BIPOC communities remain underrepresented in these jobs and disproportionately suffer from the adverse effects of human-driven climate change.

## Background

This analysis looks at access to green energy jobs (like energy efficiency, renewable energy, and green construction) by race/ethnicity, gender, education, and income in St. Paul, Minnesota, USA.

## Questions

Here are some of the questions I will explore using different datasets:

* How much climate resilience funding has St. Paul received?
* What specific green jobs are being created in St. Paul (e.g., energy efficiency, renewable energy, green construction)?
* What is the quality of these jobs? How much do they pay? What qualifications are needed (education and experience)?
* Who is getting these jobs, based on education, race/ethnicity, gender, and income levels?

## Data Sources

The data for this project comes from:

* The National Center for O\*NET Development
* 2023 Occupational Employment and Wage Survey
* Urban Institute 11 elements of job quality: Clean Energy Job Quality and Education Data
* National and local demographic data from the 2022 American Community Survey Public Use Microdata Sample (ACS PUMS)
* US Census Bureau’s 2023 QuickFacts tool
* Invest.gov
* Geocorr from the Missouri Census Data Center

I will reduce each large dataset to focus only on questions related to green jobs and job quality. Please note that some datasets have already been pre-processed in Python with specific filters applied. You can find the original raw datasets in the data folder for reference.

## Analysis

I will look at each question one by one and clean the data as I go. Some datasets might need to be combined, so I will organize the data during the analysis before exploring the results.

### Load packages and libraries

## For folder structure  
library(here)

here() starts at /Users/elhamali/Documents/Data Projects/climate-equity-workforce

library(ezknitr)  
  
## For data import/cleaning  
library(tidyverse)

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.4 ✔ readr 2.1.5  
✔ forcats 1.0.0 ✔ stringr 1.5.1  
✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
✔ purrr 1.0.2

── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()  
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(purrr)  
library(rlang)

Attaching package: 'rlang'  
  
The following objects are masked from 'package:purrr':  
  
 %@%, flatten, flatten\_chr, flatten\_dbl, flatten\_int, flatten\_lgl,  
 flatten\_raw, invoke, splice

library(forcats)  
library(readxl)  
  
## For graphing  
library(highcharter)

Registered S3 method overwritten by 'quantmod':  
 method from  
 as.zoo.data.frame zoo

library(igraph)

Attaching package: 'igraph'  
  
The following object is masked from 'package:rlang':  
  
 is\_named  
  
The following objects are masked from 'package:lubridate':  
  
 %--%, union  
  
The following objects are masked from 'package:dplyr':  
  
 as\_data\_frame, groups, union  
  
The following objects are masked from 'package:purrr':  
  
 compose, simplify  
  
The following object is masked from 'package:tidyr':  
  
 crossing  
  
The following object is masked from 'package:tibble':  
  
 as\_data\_frame  
  
The following objects are masked from 'package:stats':  
  
 decompose, spectrum  
  
The following object is masked from 'package:base':  
  
 union

library(RColorBrewer)  
library(htmlwidgets)  
library(gt)  
# library(viridis)

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

### 1. Climate Resilience Funding for St. Paul

|  |
| --- |
| RQ 1: How much climate resilience funding has the City of Saint Paul received? |
| As of June 2024, **Minnesota** received a total of $7,101,423,527 in funding for climate resilience, while **St. Paul** received $446,286,762. Specifically, as of January 2024, St. Paul has secured $433,028,012 from the Bipartisan Infrastructure Law (BIL) and $13,258,750 from the Inflation Reduction Act (IRA) for **climate resilience efforts.**  St. Paul’s funding makes up **6.28%** of Minnesota’s total climate resilience funding. Nearly **95% of St. Paul’s funding i**s allocated to **transportation** **projects**, with clean energy, buildings, and manufacturing receiving **less than 2% of the total**. It’s like filling up a swimming pool with water but using only a small 8 oz glass for clean energy, buildings, and manufacturing.  As of January 2024, St. Paul received **$8,337,843 from the BIL** and **$200,000 from the IRA** specifically for investments in clean energy, buildings, and manufacturing. |

# Import data  
funding <- read\_csv(here("processed\_data", "FundingSummary.csv"))

Warning: One or more parsing issues, call `problems()` on your data frame for details,  
e.g.:  
 dat <- vroom(...)  
 problems(dat)

Rows: 49535 Columns: 15  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (14): Agency Name, Bureau Name, Program Name, Category, Subcategory, Pro...  
dbl (1): Unique ID  
  
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.

saveRDS(funding, here("processed\_data", "funding.rds"))  
  
funding <- readRDS(here("processed\_data", "funding.rds"))

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

### Convert the `Funding Amount` to numeric and handling commas in the values  
  
funding <- funding %>%  
 mutate(`Funding Amount` = as.numeric(gsub(",", "", `Funding Amount`)))

Warning: There was 1 warning in `mutate()`.  
ℹ In argument: `Funding Amount = as.numeric(gsub(",", "", `Funding Amount`))`.  
Caused by warning:  
! NAs introduced by coercion

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

#### Filter for MN State and City of St. Paul

First, I will filter the dataset by State: **Minnesota**, and then narrow it down further to focus on the **City of St. Paul** and the surrounding region. Please note that St. Paul is part of the **Minneapolis-St. Paul-Bloomington, MN-WI** region, so I’ll ensure it’s included within that larger metropolitan area.

# Filter for Minnesota funding  
minnesota\_funding <- funding %>%  
 filter(State == "Minnesota")  
  
saveRDS(minnesota\_funding, here("processed\_data", "minnesota\_funding.rds"))

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

# Further filter for St. Paul, considering variations in city names  
st\_paul\_funding <- minnesota\_funding %>%  
 filter(str\_detect(City, regex("Saint Paul|St. Paul|South St. Paul|Minneapolis--St. Paul|Minneapolis-St. Paul", ignore\_case = TRUE)))  
  
saveRDS(st\_paul\_funding, here("processed\_data", "st\_paul\_funding.rds"))  
  
# glimpse(st\_paul\_funding)

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

#### Calculate funding for MN State and City of St. Paul

# Set options to avoid scientific notation  
options(scipen = 999)  
  
# Load Minnesota and St. Paul data  
minnesota\_funding <- readRDS(here("processed\_data", "minnesota\_funding.rds"))  
st\_paul\_funding <- readRDS(here("processed\_data", "st\_paul\_funding.rds"))  
  
# Calculate total funding for Minnesota  
total\_minnesota\_funding <- minnesota\_funding %>%  
 summarise(total\_funding = sum(`Funding Amount`, na.rm = TRUE))  
  
cat("The total amount of funding Minnesota received for climate as of June 2024 is $",   
 format(total\_minnesota\_funding$total\_funding, big.mark = ","), "\n")

The total amount of funding Minnesota received for climate as of June 2024 is $ 7,101,423,527

# Calculate total funding for St. Paul  
total\_st\_paul\_funding <- st\_paul\_funding %>%  
 summarise(total\_funding = sum(`Funding Amount`, na.rm = TRUE))  
  
cat("The total amount of funding St. Paul received for climate as of June 2024 is $",   
 format(total\_st\_paul\_funding$total\_funding, big.mark = ","), "\n")

The total amount of funding St. Paul received for climate as of June 2024 is $ 446,286,762

# Calculate total funds by funding source for St. Paul  
source\_st\_paul\_funding <- st\_paul\_funding %>%  
 group\_by(`Funding Source`) %>%  
 summarise(total\_funding = sum(`Funding Amount`, na.rm = TRUE))  
  
# Calculate specific totals for BIL and IRA  
bil\_funding <- st\_paul\_funding %>%  
 filter(`Funding Source` == "BIL") %>%  
 summarise(total\_bil = sum(`Funding Amount`, na.rm = TRUE))  
  
ira\_funding <- st\_paul\_funding %>%  
 filter(`Funding Source` == "IRA") %>%  
 summarise(total\_ira = sum(`Funding Amount`, na.rm = TRUE))  
  
# Print specific funding from BIL and IRA  
cat("As of January 2024, St. Paul has been allocated $",   
 format(bil\_funding$total\_bil, big.mark = ",", scientific = FALSE),   
 " from the Bipartisan Infrastructure Law (BIL) and $",   
 format(ira\_funding$total\_ira, big.mark = ",", scientific = FALSE),   
 " from the Inflation Reduction Act (IRA).\n")

As of January 2024, St. Paul has been allocated $ 433,028,012 from the Bipartisan Infrastructure Law (BIL) and $ 13,258,750 from the Inflation Reduction Act (IRA).

# Filter for the specific category 'Clean Energy, Buildings, and Manufacturing'  
st\_paul\_clean\_energy\_funding <- st\_paul\_funding %>%  
 filter(Category == "Clean Energy, Buildings, and Manufacturing")  
  
# Calculate total funds by funding source for the specific category  
source\_st\_paul\_clean\_energy\_funding <- st\_paul\_clean\_energy\_funding %>%  
 group\_by(`Funding Source`) %>%  
 summarise(total\_funding = sum(`Funding Amount`, na.rm = TRUE))  
  
# Calculate total funding across all sources for the specific category  
total\_st\_paul\_clean\_energy\_funding <- st\_paul\_clean\_energy\_funding %>%  
 summarise(total\_funding = sum(`Funding Amount`, na.rm = TRUE))  
  
# Calculate specific totals for BIL and IRA in the specific category  
bil\_clean\_energy\_funding <- st\_paul\_clean\_energy\_funding %>%  
 filter(`Funding Source` == "BIL") %>%  
 summarise(total\_bil = sum(`Funding Amount`, na.rm = TRUE))  
  
ira\_clean\_energy\_funding <- st\_paul\_clean\_energy\_funding %>%  
 filter(`Funding Source` == "IRA") %>%  
 summarise(total\_ira = sum(`Funding Amount`, na.rm = TRUE))  
  
# Print the total amount of funding for the specific category  
cat("The total amount of funding St. Paul received for 'Clean Energy, Buildings, and Manufacturing' as of June 2024 is $",   
 format(total\_st\_paul\_clean\_energy\_funding$total\_funding, big.mark = ","), "\n")

The total amount of funding St. Paul received for 'Clean Energy, Buildings, and Manufacturing' as of June 2024 is $ 8,537,843

# Print specific funding from BIL and IRA for the specific category  
cat("As of January 2024, St. Paul has been allocated $",   
 format(bil\_clean\_energy\_funding$total\_bil, big.mark = ",", scientific = FALSE),   
 " from the Bipartisan Infrastructure Law (BIL) and $",   
 format(ira\_clean\_energy\_funding$total\_ira, big.mark = ",", scientific = FALSE),   
 " from the Inflation Reduction Act (IRA) to invest in 'Clean Energy, Buildings, and Manufacturing'.\n")

As of January 2024, St. Paul has been allocated $ 8,337,843 from the Bipartisan Infrastructure Law (BIL) and $ 200,000 from the Inflation Reduction Act (IRA) to invest in 'Clean Energy, Buildings, and Manufacturing'.

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

As of January 2024, St. Paul has been allocated $ 433,028,012 million from the Bipartisan Infrastructure Law (BIL) and $ 13,258,750 from the Inflation Reduction Act (IRA) to invest in climate resilience efforts in total.

As of January 2024, St. Paul has been allocated $ 8,337,843 million from the Bipartisan Infrastructure Law (BIL) and $ 200,000 from the Inflation Reduction Act (IRA) to invest in ‘Clean Energy, Buildings, and Manufacturing’.

#### Calculate fraction of St. Paul’s funding from MN’s

minnesota\_funding <- readRDS(here("processed\_data", "minnesota\_funding.rds"))  
st\_paul\_funding <- readRDS(here("processed\_data", "st\_paul\_funding.rds"))  
  
# Calculate total funding for Minnesota  
total\_minnesota\_funding <- minnesota\_funding %>%  
 summarise(total\_funding = sum(`Funding Amount`, na.rm = TRUE)) %>%  
 pull(total\_funding)  
  
# Calculate total funding for St. Paul  
total\_st\_paul\_funding <- st\_paul\_funding %>%  
 summarise(total\_funding = sum(`Funding Amount`, na.rm = TRUE)) %>%  
 pull(total\_funding)  
  
# Calculate the fraction of St. Paul's funding from Minnesota's total funding  
fraction\_st\_paul <- total\_st\_paul\_funding / total\_minnesota\_funding  
  
# Output the results  
cat("The fraction of St. Paul's funding from Minnesota's total funding is: ",   
 round(fraction\_st\_paul, 4), "\n")

The fraction of St. Paul's funding from Minnesota's total funding is: 0.0628

cat("This means St. Paul's funding is", round(fraction\_st\_paul \* 100, 2), "% of Minnesota's total funding.\n")

This means St. Paul's funding is 6.28 % of Minnesota's total funding.

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

#### Visualize categories of funding for St. Paul

# Group the St. Paul data by Category and calculate the total funding for each category  
st\_paul\_category\_funding <- st\_paul\_funding %>%  
 group\_by(Category) %>%  
 summarise(total\_funding = sum(`Funding Amount`, na.rm = TRUE)) %>%  
 arrange(desc(total\_funding))  
  
colors <- brewer.pal(n = length(unique(st\_paul\_category\_funding$Category)), "Set3")  
  
# Create an interactive bar chart using highcharter  
hchart\_bar <- highchart() %>%  
 hc\_chart(type = "bar") %>%  
 hc\_xAxis(categories = st\_paul\_category\_funding$Category, title = list(text = "Category")) %>%  
 hc\_yAxis(title = list(text = "Total Funding ($)"), labels = list(format = "{value:,.0f}")) %>%  
 hc\_add\_series(name = "Total Funding",   
 data = st\_paul\_category\_funding$total\_funding,   
 colorByPoint = TRUE,   
 colors = colors) %>%  
 hc\_title(text = "Total Funding by Category in St. Paul") %>%  
 hc\_tooltip(pointFormat = "Total Funding: ${point.y:,.0f}") %>%  
 hc\_exporting(  
 enabled = TRUE,  
 buttons = list(contextButton = list(menuItems = c("downloadPNG", "downloadJPEG", "downloadSVG", "downloadPDF")))  
 )  
  
# Saving the chart as an HTML file  
saveWidget(hchart\_bar, file = here("graphs", "st\_paul\_funding\_bar.html"))

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

A quick glance tells us that almost **95%** of St. Paul’s funding goes to transportation efforts. Clean energy, buildings and manufacturing received less than **2%** of funding.

# Create an interactive pie chart using highcharter  
hchart\_pie <- highchart() %>%  
 hc\_chart(type = "pie") %>%  
 hc\_add\_series(name = "Total Funding",   
 data = list\_parse2(st\_paul\_category\_funding %>%   
 mutate(name = Category, y = total\_funding)),   
 colors = colors) %>%  
 hc\_title(text = "Total Funding by Category in St. Paul") %>%  
 hc\_tooltip(pointFormat = "Total Funding: ${point.y:,.0f}") %>%  
 hc\_plotOptions(pie = list(innerSize = '50%', dataLabels = list(enabled = TRUE))) %>%  
 hc\_exporting(  
 enabled = TRUE,  
 buttons = list(contextButton = list(menuItems = c("downloadPNG", "downloadJPEG", "downloadSVG", "downloadPDF")))  
 )  
  
saveWidget(hchart\_pie, file = here("graphs", "st\_paul\_funding\_pie.html"))

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

## Export the funding data to CSV for graphing  
write.csv(minnesota\_funding, here("processed\_data", "minnesota\_funding.csv"), row.names = FALSE)  
write.csv(st\_paul\_funding, here("processed\_data", "st\_paul\_funding.csv"), row.names = FALSE)

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

### 2. Types of Green Jobs in St. Paul

|  |
| --- |
| RQ 2: What specific green jobs are being created in the Minneapolis-Saint Paul metropolitan area and nationally (e.g., energy efficiency, renewable energy, green construction)? |
| Nationally  There’s a total of **17,119,730 employed people** in green jobs nationally. Specifically, in **Energy Efficiency**, there are 4,928,520 (28.79 %), in **Green Construction** there are 10,624,140 (62.06 %), and in **Renewable Energy Generation** there are 1,567,070 (9.15 %).  The **mean annual wage** for the occupation in U.S. dollars for green jobs is $78,363.4, and for non-green jobs is $73,763.67. That means green jobs pay **$4,599.73 more** than non-green jobs **nationally**.  The **mean hourly wage** for the occupation in U.S. dollars for green jobs is $37.67547, and for non-green jobs is $34.80. That means green jobs pay **$2.88 more** than non-green jobs **nationally**.  Minneapolis-Saint Paul Metropolitan Area  There’s a total of **214,340 employed people** in green jobs in the Minneapolis-Saint Paul metropolitan area. Specifically, in **Energy Efficiency**, there are 66,410 ( 30.98 %), in **Green Construction** there are 124,680 ( 58.17 %), and in **Renewable Energy Generation** there are 23,250 ( 10.85 %).  The **mean annual wage** for the occupation in U.S. dollars for green jobs **in this area** is $84,561.7, and for non-green jobs is $77,192.53. That means green jobs in Saint Paul pay $7,369.169 more than non-green jobs in this area.  The **mean hourly wage** for the occupation in U.S. dollars for green jobs **in this area** is $40.65, and for non-green jobs is $36.31. That means green jobs in Saint Paul pay **$4.35 more** than non-green jobs in this area. |

#### Green jobs nationally

# Import national jobs data  
national\_jobs <- read\_csv(here("processed\_data", "OWES\_and\_ONET-National.csv"))

Rows: 1420 Columns: 34  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (21): AREA\_TITLE, PRIM\_STATE, NAICS\_TITLE, I\_GROUP, OCC\_CODE, OCC\_TITLE,...  
dbl (7): AREA, AREA\_TYPE, NAICS, OWN\_CODE, TOT\_EMP, EMP\_PRSE, MEAN\_PRSE  
lgl (6): JOBS\_1000, LOC\_QUOTIENT, PCT\_TOTAL, PCT\_RPT, ANNUAL, HOURLY  
  
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.

saveRDS(national\_jobs, here("processed\_data", "national\_jobs.rds"))  
  
national\_jobs <- readRDS(here("processed\_data", "national\_jobs.rds"))

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

Here, we’d want to filter to only green jobs

# Convert necessary columns to numeric where needed  
national\_jobs <- national\_jobs %>%  
 mutate(  
 TOT\_EMP = as.numeric(TOT\_EMP),  
 # JOBS\_1000 = as.numeric(JOBS\_1000),  
 # PCT\_TOTAL = as.numeric(PCT\_TOTAL),  
 H\_MEAN = as.numeric(H\_MEAN),  
 A\_MEAN = as.numeric(A\_MEAN),  
 A\_MEDIAN = as.numeric(A\_MEDIAN),  
 H\_MEDIAN = as.numeric(H\_MEDIAN)  
 )

Warning: There were 4 warnings in `mutate()`.  
The first warning was:  
ℹ In argument: `H\_MEAN = as.numeric(H\_MEAN)`.  
Caused by warning:  
! NAs introduced by coercion  
ℹ Run `dplyr::last\_dplyr\_warnings()` to see the 3 remaining warnings.

# Filter the dataset to include only relevant sectors  
filtered\_jobs <- national\_jobs %>%  
 filter(`O\*NET-SOC Sector` %in% c("Energy Efficiency", "Renewable Energy Generation", "Green Construction"))  
  
# Function to summarize data for each sector  
summarize\_by\_sector <- function(df) {  
 df %>%  
 summarize(  
 TOT\_EMP = sum(TOT\_EMP, na.rm = TRUE),  
 # JOBS\_1000 = sum(JOBS\_1000 \* TOT\_EMP, na.rm = TRUE) / sum(TOT\_EMP, na.rm = TRUE),  
 # PCT\_TOTAL = sum(PCT\_TOTAL \* TOT\_EMP, na.rm = TRUE) / sum(TOT\_EMP, na.rm = TRUE),  
 H\_MEAN = mean(H\_MEAN, na.rm = TRUE),  
 A\_MEAN = mean(A\_MEAN, na.rm = TRUE),  
 A\_MEDIAN = median(A\_MEDIAN, na.rm = TRUE),  
 H\_MEDIAN = median(H\_MEDIAN, na.rm = TRUE)  
 )  
}  
  
# Summarize the data for each sector and overall  
sector\_summary <- filtered\_jobs %>%  
 group\_by(`O\*NET-SOC Sector`) %>%  
 summarize\_by\_sector()  
  
# Calculate the summary for all sectors combined  
overall\_summary <- filtered\_jobs %>%  
 summarize\_by\_sector()  
  
# Combine the results: sector-wise and overall  
final\_summary <- bind\_rows(sector\_summary, tibble(`O\*NET-SOC Sector` = "All", overall\_summary))  
  
# Save the final summary as an RDS file and CSV for future reference  
saveRDS(final\_summary, here("processed\_data", "sector\_summary.rds"))  
write\_csv(final\_summary, here("processed\_data", "sector\_summary.csv"))  
  
# Output the final summary to the user  
print(final\_summary)

# A tibble: 4 × 6  
 `O\*NET-SOC Sector` TOT\_EMP H\_MEAN A\_MEAN A\_MEDIAN H\_MEDIAN  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 Energy Efficiency 4928520 43.0 89371 86355 41.5  
2 Green Construction 10624140 33.9 70506. 60165 28.9  
3 Renewable Energy Generation 1567070 42.3 88028. 97010 46.6  
4 All 17119730 37.7 78363. 67640 32.5

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

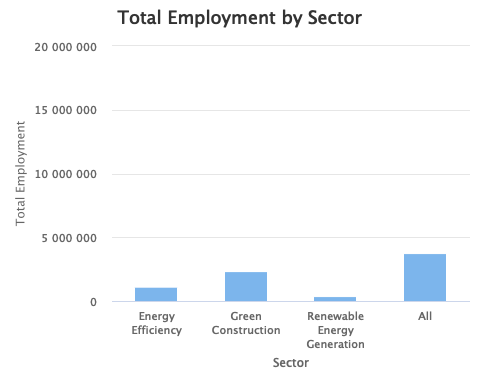
# Calculate total employment and sector percentages  
total\_green\_jobs <- final\_summary %>% filter(`O\*NET-SOC Sector` == "All") %>% pull(TOT\_EMP)  
  
energy\_efficiency\_jobs <- final\_summary %>% filter(`O\*NET-SOC Sector` == "Energy Efficiency") %>% pull(TOT\_EMP)  
green\_construction\_jobs <- final\_summary %>% filter(`O\*NET-SOC Sector` == "Green Construction") %>% pull(TOT\_EMP)  
renewable\_energy\_jobs <- final\_summary %>% filter(`O\*NET-SOC Sector` == "Renewable Energy Generation") %>% pull(TOT\_EMP)  
  
# Calculate the percentages  
energy\_efficiency\_pct <- round((energy\_efficiency\_jobs / total\_green\_jobs) \* 100, 2)  
green\_construction\_pct <- round((green\_construction\_jobs / total\_green\_jobs) \* 100, 2)  
renewable\_energy\_pct <- round((renewable\_energy\_jobs / total\_green\_jobs) \* 100, 2)  
  
# Create the concatenated sentence  
cat("There's a total of", format(total\_green\_jobs, big.mark = ",", scientific = FALSE),   
 "employed people in green jobs nationally. Specifically, in Energy Efficiency, there are",   
 format(energy\_efficiency\_jobs, big.mark = ",", scientific = FALSE),   
 "(", energy\_efficiency\_pct, "%), in Green Construction there are",   
 format(green\_construction\_jobs, big.mark = ",", scientific = FALSE),   
 "(", green\_construction\_pct, "%), and in Renewable Energy Generation there are",   
 format(renewable\_energy\_jobs, big.mark = ",", scientific = FALSE),   
 "(", renewable\_energy\_pct, "%).\n")

There's a total of 17,119,730 employed people in green jobs nationally. Specifically, in Energy Efficiency, there are 4,928,520 ( 28.79 %), in Green Construction there are 10,624,140 ( 62.06 %), and in Renewable Energy Generation there are 1,567,070 ( 9.15 %).

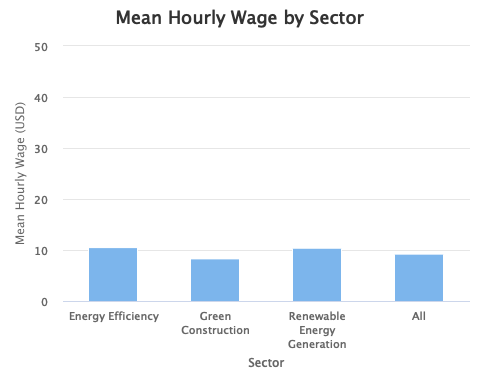
Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

Let’s visualize this so it’s easier to compare across all green sectors

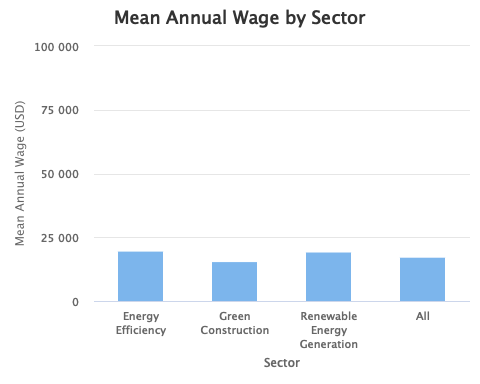
# Convert the O\*NET-SOC Sector to a factor for ordering in the chart  
final\_summary <- final\_summary %>%  
 mutate(`O\*NET-SOC Sector` = factor(`O\*NET-SOC Sector`, levels = c("Energy Efficiency", "Green Construction", "Renewable Energy Generation", "All")))  
  
# Visualizing TOT\_EMP across the sectors  
hchart(final\_summary, "column", hcaes(x = `O\*NET-SOC Sector`, y = TOT\_EMP)) %>%  
 hc\_title(text = "Total Employment by Sector") %>%  
 hc\_xAxis(title = list(text = "Sector")) %>%  
 hc\_yAxis(title = list(text = "Total Employment"), labels = list(format = "{value:,0f}")) %>%  
 hc\_tooltip(pointFormat = '<b>{point.y:,0f}</b>')



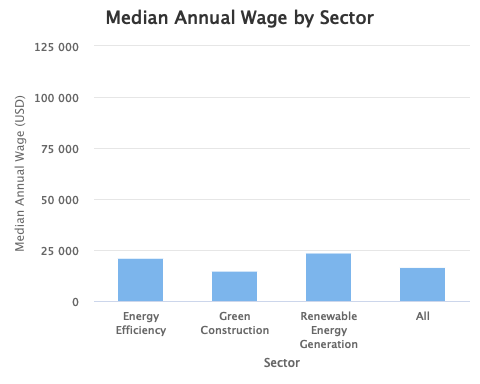
# Visualizing H\_MEAN (Mean Hourly Wage) across the sectors  
hchart(final\_summary, "column", hcaes(x = `O\*NET-SOC Sector`, y = H\_MEAN)) %>%  
 hc\_title(text = "Mean Hourly Wage by Sector") %>%  
 hc\_xAxis(title = list(text = "Sector")) %>%  
 hc\_yAxis(title = list(text = "Mean Hourly Wage (USD)")) %>%  
 hc\_tooltip(pointFormat = '<b>{point.y:.2f} USD</b>')



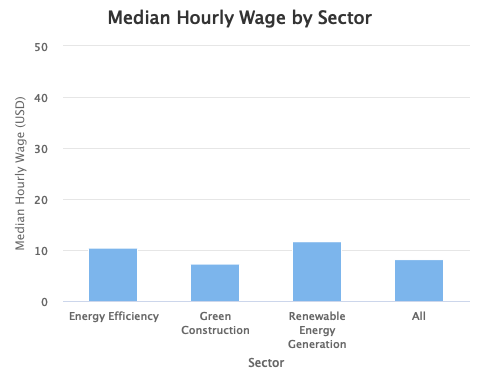
# Visualizing A\_MEAN (Mean Annual Wage) across the sectors  
hchart(final\_summary, "column", hcaes(x = `O\*NET-SOC Sector`, y = A\_MEAN)) %>%  
 hc\_title(text = "Mean Annual Wage by Sector") %>%  
 hc\_xAxis(title = list(text = "Sector")) %>%  
 hc\_yAxis(title = list(text = "Mean Annual Wage (USD)"), labels = list(format = "{value:,0f}")) %>%  
 hc\_tooltip(pointFormat = '<b>{point.y:,0f} USD</b>')



# Visualizing A\_MEDIAN (Median Annual Wage) across the sectors  
hchart(final\_summary, "column", hcaes(x = `O\*NET-SOC Sector`, y = A\_MEDIAN)) %>%  
 hc\_title(text = "Median Annual Wage by Sector") %>%  
 hc\_xAxis(title = list(text = "Sector")) %>%  
 hc\_yAxis(title = list(text = "Median Annual Wage (USD)"), labels = list(format = "{value:,0f}")) %>%  
 hc\_tooltip(pointFormat = '<b>{point.y:,0f} USD</b>')



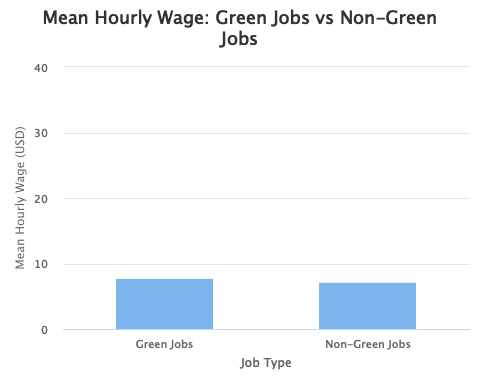
# Visualizing H\_MEDIAN (Median Hourly Wage) across the sectors  
hchart(final\_summary, "column", hcaes(x = `O\*NET-SOC Sector`, y = H\_MEDIAN)) %>%  
 hc\_title(text = "Median Hourly Wage by Sector") %>%  
 hc\_xAxis(title = list(text = "Sector")) %>%  
 hc\_yAxis(title = list(text = "Median Hourly Wage (USD)")) %>%  
 hc\_tooltip(pointFormat = '<b>{point.y:.2f} USD</b>')



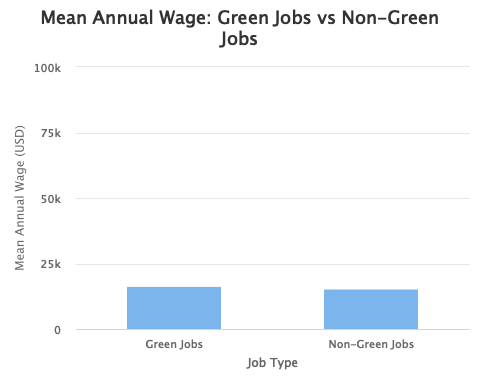
Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

I’m also curious about the differences between green jobs and non-green jobs for mean hourly wage and mean annual wage.

# Define green jobs as sectors related to energy and construction  
green\_jobs\_sectors <- c("Energy Efficiency", "Renewable Energy Generation", "Green Construction")  
  
# Add a new column to identify green and non-green jobs  
national\_jobs <- national\_jobs %>%  
 mutate(  
 Job\_Type = ifelse(`O\*NET-SOC Sector` %in% green\_jobs\_sectors, "Green Jobs", "Non-Green Jobs")  
 )  
  
# Group by job type (Green vs Non-Green) and calculate mean wages  
job\_type\_summary <- national\_jobs %>%  
 group\_by(Job\_Type) %>%  
 summarize(  
 H\_MEAN = mean(H\_MEAN, na.rm = TRUE),  
 A\_MEAN = mean(A\_MEAN, na.rm = TRUE)  
 )  
  
# Visualizing Mean Hourly Wage (H\_MEAN) for Green vs Non-Green Jobs  
hchart(job\_type\_summary, "column", hcaes(x = Job\_Type, y = H\_MEAN)) %>%  
 hc\_title(text = "Mean Hourly Wage: Green Jobs vs Non-Green Jobs") %>%  
 hc\_xAxis(title = list(text = "Job Type")) %>%  
 hc\_yAxis(title = list(text = "Mean Hourly Wage (USD)")) %>%  
 hc\_tooltip(pointFormat = '<b>{point.y:.2f} USD</b>')



# Visualizing Mean Annual Wage (A\_MEAN) for Green vs Non-Green Jobs  
hchart(job\_type\_summary, "column", hcaes(x = Job\_Type, y = A\_MEAN)) %>%  
 hc\_title(text = "Mean Annual Wage: Green Jobs vs Non-Green Jobs") %>%  
 hc\_xAxis(title = list(text = "Job Type")) %>%  
 hc\_yAxis(title = list(text = "Mean Annual Wage (USD)")) %>%  
 hc\_tooltip(pointFormat = '<b>{point.y:,0f} USD</b>')



Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

# Summarizing core findings nationally  
  
# Extract green and non-green job wage data  
green\_wages <- job\_type\_summary %>% filter(Job\_Type == "Green Jobs")  
non\_green\_wages <- job\_type\_summary %>% filter(Job\_Type == "Non-Green Jobs")  
  
# Calculate the difference between green and non-green jobs  
difference\_annual <- green\_wages$A\_MEAN - non\_green\_wages$A\_MEAN  
difference\_hourly <- green\_wages$H\_MEAN - non\_green\_wages$H\_MEAN  
  
# Format and print the sentences  
cat("The mean annual wage for the occupation in U.S. dollars for green jobs is $",   
 format(green\_wages$A\_MEAN, big.mark = ",", scientific = FALSE),   
 ", and for non-green jobs is $",   
 format(non\_green\_wages$A\_MEAN, big.mark = ",", scientific = FALSE),   
 ". That means green jobs pay $",   
 format(abs(difference\_annual), big.mark = ",", scientific = FALSE),   
 ifelse(difference\_annual > 0, " more", " less"),   
 " than non-green jobs nationally.\n", sep = "")

The mean annual wage for the occupation in U.S. dollars for green jobs is $78,363.4, and for non-green jobs is $73,763.67. That means green jobs pay $4,599.726 more than non-green jobs nationally.

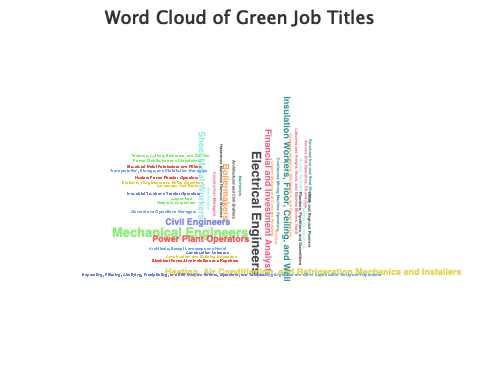
cat("The mean hourly wage for the occupation in U.S. dollars for green jobs is $",   
 format(green\_wages$H\_MEAN, big.mark = ",", scientific = FALSE),   
 ", and for non-green jobs is $",   
 format(non\_green\_wages$H\_MEAN, big.mark = ",", scientific = FALSE),   
 ". That means green jobs pay $",   
 format(abs(difference\_hourly), big.mark = ",", scientific = FALSE),   
 ifelse(difference\_hourly > 0, " more", " less"),   
 " than non-green jobs nationally.\n", sep = "")

The mean hourly wage for the occupation in U.S. dollars for green jobs is $37.67547, and for non-green jobs is $34.79641. That means green jobs pay $2.879063 more than non-green jobs nationally.

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

I’d like to see a word cloud of different job titles for each sector

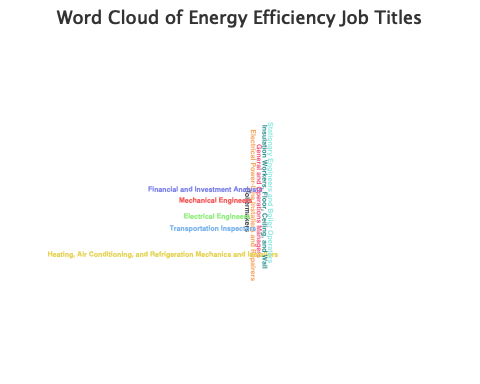
# Filter the dataset for green jobs only  
green\_jobs <- national\_jobs %>%  
 filter(`O\*NET-SOC Sector` %in% c("Energy Efficiency", "Renewable Energy Generation", "Green Construction"))  
  
# Extract job titles and count their occurrences  
job\_titles <- green\_jobs %>%  
 count(OCC\_TITLE, sort = TRUE)  
  
# Create a word cloud using highcharter  
hchart(  
 job\_titles,   
 "wordcloud",   
 hcaes(name = OCC\_TITLE, weight = n)  
) %>%  
 hc\_title(text = "Word Cloud of Green Job Titles")



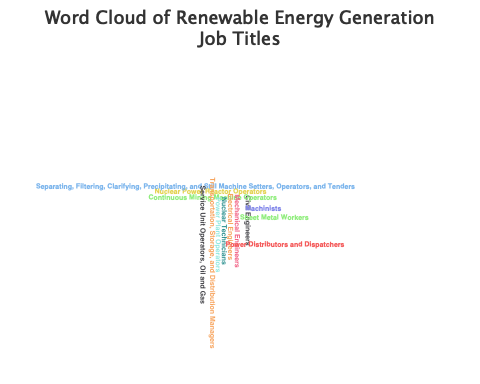
Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

Now let’s create separate word clouds for each of the green sectors (“Energy Efficiency”, “Renewable Energy Generation”, and “Green Construction”).

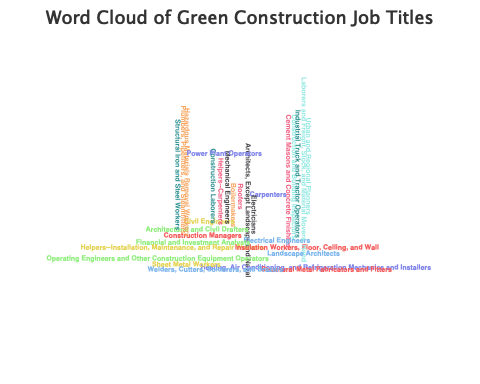
# Filter the dataset for each sector  
energy\_efficiency\_jobs <- national\_jobs %>%  
 filter(`O\*NET-SOC Sector` == "Energy Efficiency")  
  
renewable\_energy\_jobs <- national\_jobs %>%  
 filter(`O\*NET-SOC Sector` == "Renewable Energy Generation")  
  
green\_construction\_jobs <- national\_jobs %>%  
 filter(`O\*NET-SOC Sector` == "Green Construction")  
  
# Create a function to generate word clouds  
generate\_wordcloud <- function(data, sector\_name) {  
 job\_titles <- data %>%  
 count(OCC\_TITLE, sort = TRUE)  
   
 hchart(  
 job\_titles,   
 "wordcloud",   
 hcaes(name = OCC\_TITLE, weight = n)  
 ) %>%  
 hc\_title(text = paste("Word Cloud of", sector\_name, "Job Titles"))  
}  
  
# Generate word cloud for Energy Efficiency  
energy\_efficiency\_wordcloud <- generate\_wordcloud(energy\_efficiency\_jobs, "Energy Efficiency")  
  
# Generate word cloud for Renewable Energy Generation  
renewable\_energy\_wordcloud <- generate\_wordcloud(renewable\_energy\_jobs, "Renewable Energy Generation")  
  
# Generate word cloud for Green Construction  
green\_construction\_wordcloud <- generate\_wordcloud(green\_construction\_jobs, "Green Construction")  
  
# Display the word clouds  
energy\_efficiency\_wordcloud



renewable\_energy\_wordcloud



green\_construction\_wordcloud



Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

Let’s export for graphing

## Export the national jobs data to CSV for graphing  
write.csv(national\_jobs, here("processed\_data", "national\_jobs.csv"), row.names = FALSE)  
  
# Export the job\_type\_summary dataset to CSV for graphing  
write.csv(job\_type\_summary, here("processed\_data", "national\_job\_type\_summary.csv"), row.names = FALSE)

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

#### Green jobs in St. Paul

# Import St. Paul jobs data  
st\_paul\_jobs <- read\_csv(here("processed\_data", "OWES\_and\_ONET-St\_Paul.csv"))

Rows: 742 Columns: 34  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (26): AREA\_TITLE, PRIM\_STATE, NAICS\_TITLE, I\_GROUP, OCC\_CODE, OCC\_TITLE,...  
dbl (4): AREA, AREA\_TYPE, NAICS, OWN\_CODE  
lgl (4): PCT\_TOTAL, PCT\_RPT, ANNUAL, HOURLY  
  
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.

saveRDS(st\_paul\_jobs, here("processed\_data", "st\_paul\_jobs.rds"))  
  
st\_paul\_jobs <- readRDS(here("processed\_data", "st\_paul\_jobs.rds"))

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

# Convert necessary columns to numeric where needed  
st\_paul\_jobs <- st\_paul\_jobs %>%  
 mutate(  
 TOT\_EMP = as.numeric(TOT\_EMP),  
 H\_MEAN = as.numeric(H\_MEAN),  
 A\_MEAN = as.numeric(A\_MEAN),  
 A\_MEDIAN = as.numeric(A\_MEDIAN),  
 H\_MEDIAN = as.numeric(H\_MEDIAN)  
 )

Warning: There were 5 warnings in `mutate()`.  
The first warning was:  
ℹ In argument: `TOT\_EMP = as.numeric(TOT\_EMP)`.  
Caused by warning:  
! NAs introduced by coercion  
ℹ Run `dplyr::last\_dplyr\_warnings()` to see the 4 remaining warnings.

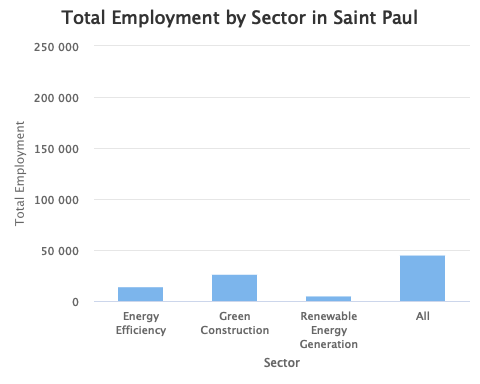
# Filter the dataset to include only relevant green sectors  
filtered\_st\_paul\_jobs <- st\_paul\_jobs %>%  
 filter(`O\*NET-SOC Sector` %in% c("Energy Efficiency", "Renewable Energy Generation", "Green Construction"))  
  
# Function to summarize data for each sector  
summarize\_by\_sector <- function(df) {  
 df %>%  
 summarize(  
 TOT\_EMP = sum(TOT\_EMP, na.rm = TRUE),  
 H\_MEAN = mean(H\_MEAN, na.rm = TRUE),  
 A\_MEAN = mean(A\_MEAN, na.rm = TRUE),  
 A\_MEDIAN = median(A\_MEDIAN, na.rm = TRUE),  
 H\_MEDIAN = median(H\_MEDIAN, na.rm = TRUE)  
 )  
}  
  
# Summarize the data for each sector and overall  
sector\_summary\_st\_paul <- filtered\_st\_paul\_jobs %>%  
 group\_by(`O\*NET-SOC Sector`) %>%  
 summarize\_by\_sector()  
  
# Calculate the summary for all sectors combined  
overall\_summary\_st\_paul <- filtered\_st\_paul\_jobs %>%  
 summarize\_by\_sector()  
  
# Combine the results: sector-wise and overall  
final\_summary\_st\_paul <- bind\_rows(sector\_summary\_st\_paul, tibble(`O\*NET-SOC Sector` = "All", overall\_summary\_st\_paul))  
  
# Save the final summary as an RDS file and CSV for future reference  
saveRDS(final\_summary\_st\_paul, here("processed\_data", "sector\_summary\_st\_paul.rds"))  
write\_csv(final\_summary\_st\_paul, here("processed\_data", "sector\_summary\_st\_paul.csv"))  
  
# Output the final summary to the user  
print(final\_summary\_st\_paul)

# A tibble: 4 × 6  
 `O\*NET-SOC Sector` TOT\_EMP H\_MEAN A\_MEAN A\_MEDIAN H\_MEDIAN  
 <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 Energy Efficiency 66410 45.5 94669. 98740 47.5  
2 Green Construction 124680 37.9 78809. 75800 36.4  
3 Renewable Energy Generation 23250 44.7 92991. 99690 47.9  
4 All 214340 40.7 84562. 82170 39.5

# Calculate total employment and sector percentages for St. Paul  
total\_green\_jobs\_st\_paul <- final\_summary\_st\_paul %>% filter(`O\*NET-SOC Sector` == "All") %>% pull(TOT\_EMP)  
  
energy\_efficiency\_jobs\_st\_paul <- final\_summary\_st\_paul %>% filter(`O\*NET-SOC Sector` == "Energy Efficiency") %>% pull(TOT\_EMP)  
green\_construction\_jobs\_st\_paul <- final\_summary\_st\_paul %>% filter(`O\*NET-SOC Sector` == "Green Construction") %>% pull(TOT\_EMP)  
renewable\_energy\_jobs\_st\_paul <- final\_summary\_st\_paul %>% filter(`O\*NET-SOC Sector` == "Renewable Energy Generation") %>% pull(TOT\_EMP)  
  
# Calculate the percentages  
energy\_efficiency\_pct\_st\_paul <- round((energy\_efficiency\_jobs\_st\_paul / total\_green\_jobs\_st\_paul) \* 100, 2)  
green\_construction\_pct\_st\_paul <- round((green\_construction\_jobs\_st\_paul / total\_green\_jobs\_st\_paul) \* 100, 2)  
renewable\_energy\_pct\_st\_paul <- round((renewable\_energy\_jobs\_st\_paul / total\_green\_jobs\_st\_paul) \* 100, 2)  
  
# Create the concatenated sentence for St. Paul  
cat("There's a total of", format(total\_green\_jobs\_st\_paul, big.mark = ",", scientific = FALSE),   
 "employed people in green jobs in Saint Paul. Specifically, in Energy Efficiency, there are",   
 format(energy\_efficiency\_jobs\_st\_paul, big.mark = ",", scientific = FALSE),   
 "(", energy\_efficiency\_pct\_st\_paul, "%), in Green Construction there are",   
 format(green\_construction\_jobs\_st\_paul, big.mark = ",", scientific = FALSE),   
 "(", green\_construction\_pct\_st\_paul, "%), and in Renewable Energy Generation there are",   
 format(renewable\_energy\_jobs\_st\_paul, big.mark = ",", scientific = FALSE),   
 "(", renewable\_energy\_pct\_st\_paul, "%).\n")

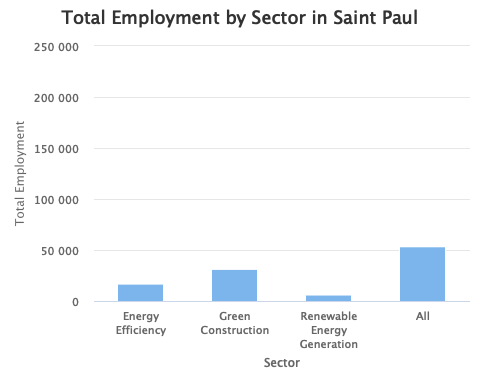
There's a total of 214,340 employed people in green jobs in Saint Paul. Specifically, in Energy Efficiency, there are 66,410 ( 30.98 %), in Green Construction there are 124,680 ( 58.17 %), and in Renewable Energy Generation there are 23,250 ( 10.85 %).

# Visualizing TOT\_EMP across the green sectors for St. Paul  
final\_summary\_st\_paul <- final\_summary\_st\_paul %>%  
 mutate(`O\*NET-SOC Sector` = factor(`O\*NET-SOC Sector`, levels = c("Energy Efficiency", "Green Construction", "Renewable Energy Generation", "All")))  
  
hchart(final\_summary\_st\_paul, "column", hcaes(x = `O\*NET-SOC Sector`, y = TOT\_EMP)) %>%  
 hc\_title(text = "Total Employment by Sector in Saint Paul") %>%  
 hc\_xAxis(title = list(text = "Sector")) %>%  
 hc\_yAxis(title = list(text = "Total Employment"), labels = list(format = "{value:,0f}")) %>%  
 hc\_tooltip(pointFormat = '<b>{point.y:,0f}</b>')

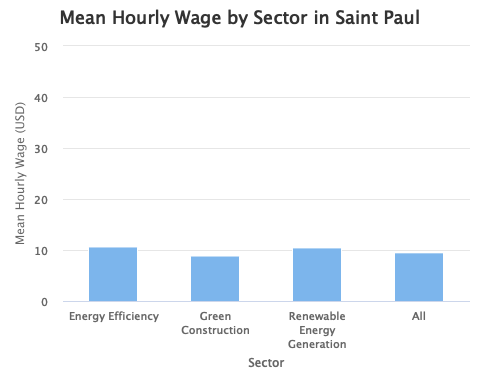


Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

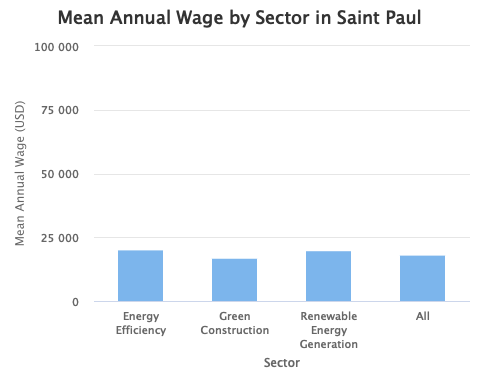
# Convert the O\*NET-SOC Sector to a factor for ordering in the chart for St. Paul  
final\_summary\_st\_paul <- final\_summary\_st\_paul %>%  
 mutate(`O\*NET-SOC Sector` = factor(`O\*NET-SOC Sector`, levels = c("Energy Efficiency", "Green Construction", "Renewable Energy Generation", "All")))  
  
# Visualizing TOT\_EMP across the sectors for St. Paul  
hchart(final\_summary\_st\_paul, "column", hcaes(x = `O\*NET-SOC Sector`, y = TOT\_EMP)) %>%  
 hc\_title(text = "Total Employment by Sector in Saint Paul") %>%  
 hc\_xAxis(title = list(text = "Sector")) %>%  
 hc\_yAxis(title = list(text = "Total Employment"), labels = list(format = "{value:,0f}")) %>%  
 hc\_tooltip(pointFormat = '<b>{point.y:,0f}</b>')



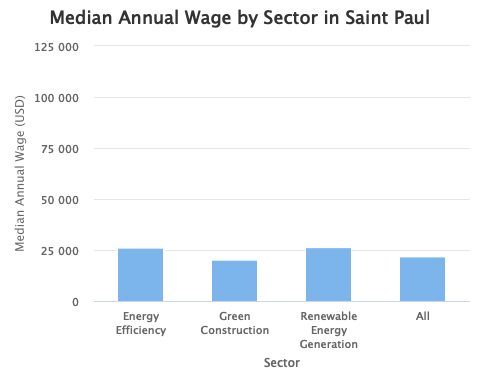
# Visualizing H\_MEAN (Mean Hourly Wage) across the sectors for St. Paul  
hchart(final\_summary\_st\_paul, "column", hcaes(x = `O\*NET-SOC Sector`, y = H\_MEAN)) %>%  
 hc\_title(text = "Mean Hourly Wage by Sector in Saint Paul") %>%  
 hc\_xAxis(title = list(text = "Sector")) %>%  
 hc\_yAxis(title = list(text = "Mean Hourly Wage (USD)")) %>%  
 hc\_tooltip(pointFormat = '<b>{point.y:.2f} USD</b>')



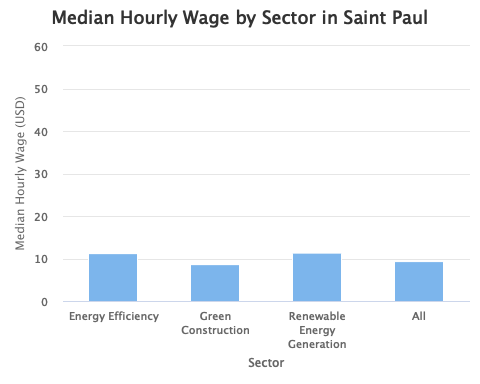
# Visualizing A\_MEAN (Mean Annual Wage) across the sectors for St. Paul  
hchart(final\_summary\_st\_paul, "column", hcaes(x = `O\*NET-SOC Sector`, y = A\_MEAN)) %>%  
 hc\_title(text = "Mean Annual Wage by Sector in Saint Paul") %>%  
 hc\_xAxis(title = list(text = "Sector")) %>%  
 hc\_yAxis(title = list(text = "Mean Annual Wage (USD)"), labels = list(format = "{value:,0f}")) %>%  
 hc\_tooltip(pointFormat = '<b>{point.y:,0f} USD</b>')



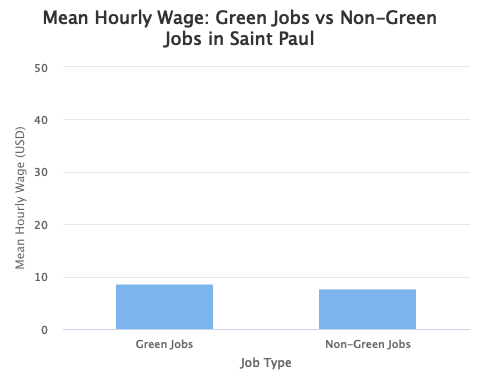
# Visualizing A\_MEDIAN (Median Annual Wage) across the sectors for St. Paul  
hchart(final\_summary\_st\_paul, "column", hcaes(x = `O\*NET-SOC Sector`, y = A\_MEDIAN)) %>%  
 hc\_title(text = "Median Annual Wage by Sector in Saint Paul") %>%  
 hc\_xAxis(title = list(text = "Sector")) %>%  
 hc\_yAxis(title = list(text = "Median Annual Wage (USD)"), labels = list(format = "{value:,0f}")) %>%  
 hc\_tooltip(pointFormat = '<b>{point.y:,0f} USD</b>')



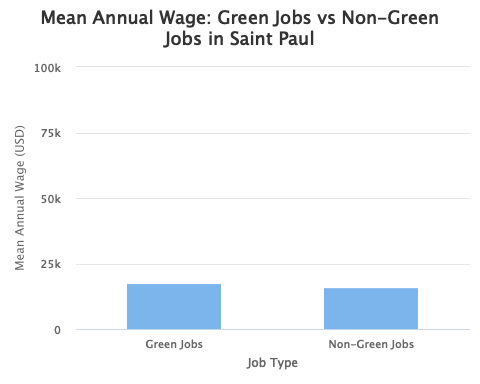
# Visualizing H\_MEDIAN (Median Hourly Wage) across the sectors for St. Paul  
hchart(final\_summary\_st\_paul, "column", hcaes(x = `O\*NET-SOC Sector`, y = H\_MEDIAN)) %>%  
 hc\_title(text = "Median Hourly Wage by Sector in Saint Paul") %>%  
 hc\_xAxis(title = list(text = "Sector")) %>%  
 hc\_yAxis(title = list(text = "Median Hourly Wage (USD)")) %>%  
 hc\_tooltip(pointFormat = '<b>{point.y:.2f} USD</b>')



# Define green jobs as sectors related to energy and construction for St. Paul  
green\_jobs\_sectors\_st\_paul <- c("Energy Efficiency", "Renewable Energy Generation", "Green Construction")  
  
# Add a new column to identify green and non-green jobs for St. Paul  
st\_paul\_jobs <- st\_paul\_jobs %>%  
 mutate(  
 Job\_Type = ifelse(`O\*NET-SOC Sector` %in% green\_jobs\_sectors\_st\_paul, "Green Jobs", "Non-Green Jobs")  
 )  
  
# Group by job type (Green vs Non-Green) and calculate mean wages for St. Paul  
job\_type\_summary\_st\_paul <- st\_paul\_jobs %>%  
 group\_by(Job\_Type) %>%  
 summarize(  
 H\_MEAN = mean(H\_MEAN, na.rm = TRUE),  
 A\_MEAN = mean(A\_MEAN, na.rm = TRUE)  
 )  
  
# Visualizing Mean Hourly Wage (H\_MEAN) for Green vs Non-Green Jobs in St. Paul  
hchart(job\_type\_summary\_st\_paul, "column", hcaes(x = Job\_Type, y = H\_MEAN)) %>%  
 hc\_title(text = "Mean Hourly Wage: Green Jobs vs Non-Green Jobs in Saint Paul") %>%  
 hc\_xAxis(title = list(text = "Job Type")) %>%  
 hc\_yAxis(title = list(text = "Mean Hourly Wage (USD)")) %>%  
 hc\_tooltip(pointFormat = '<b>{point.y:.2f} USD</b>')



# Visualizing Mean Annual Wage (A\_MEAN) for Green vs Non-Green Jobs in St. Paul  
hchart(job\_type\_summary\_st\_paul, "column", hcaes(x = Job\_Type, y = A\_MEAN)) %>%  
 hc\_title(text = "Mean Annual Wage: Green Jobs vs Non-Green Jobs in Saint Paul") %>%  
 hc\_xAxis(title = list(text = "Job Type")) %>%  
 hc\_yAxis(title = list(text = "Mean Annual Wage (USD)")) %>%  
 hc\_tooltip(pointFormat = '<b>{point.y:,0f} USD</b>')



# Summarizing core findings for Saint Paul  
  
# Extract green and non-green job wage data for St. Paul  
green\_wages\_st\_paul <- job\_type\_summary\_st\_paul %>% filter(Job\_Type == "Green Jobs")  
non\_green\_wages\_st\_paul <- job\_type\_summary\_st\_paul %>% filter(Job\_Type == "Non-Green Jobs")  
  
# Calculate the difference between green and non-green jobs for St. Paul  
difference\_annual\_st\_paul <- green\_wages\_st\_paul$A\_MEAN - non\_green\_wages\_st\_paul$A\_MEAN  
difference\_hourly\_st\_paul <- green\_wages\_st\_paul$H\_MEAN - non\_green\_wages\_st\_paul$H\_MEAN  
  
# Format and print the sentences for Saint Paul  
cat("The mean annual wage for the occupation in U.S. dollars for green jobs in Saint Paul is $",   
 format(green\_wages\_st\_paul$A\_MEAN, big.mark = ",", scientific = FALSE),   
 ", and for non-green jobs is $",   
 format(non\_green\_wages\_st\_paul$A\_MEAN, big.mark = ",", scientific = FALSE),   
 ". That means green jobs in Saint Paul pay $",   
 format(abs(difference\_annual\_st\_paul), big.mark = ",", scientific = FALSE),   
 ifelse(difference\_annual\_st\_paul > 0, " more", " less"),   
 " than non-green jobs in Saint Paul.\n", sep = "")

The mean annual wage for the occupation in U.S. dollars for green jobs in Saint Paul is $84,561.7, and for non-green jobs is $77,192.53. That means green jobs in Saint Paul pay $7,369.169 more than non-green jobs in Saint Paul.

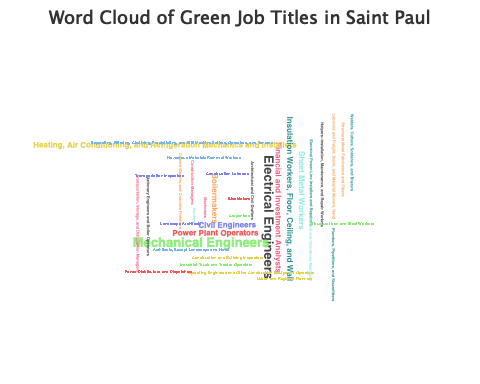
cat("The mean hourly wage for the occupation in U.S. dollars for green jobs in Saint Paul is $",   
 format(green\_wages\_st\_paul$H\_MEAN, big.mark = ",", scientific = FALSE),   
 ", and for non-green jobs is $",   
 format(non\_green\_wages\_st\_paul$H\_MEAN, big.mark = ",", scientific = FALSE),   
 ". That means green jobs in Saint Paul pay $",   
 format(abs(difference\_hourly\_st\_paul), big.mark = ",", scientific = FALSE),   
 ifelse(difference\_hourly\_st\_paul > 0, " more", " less"),   
 " than non-green jobs in Saint Paul.\n", sep = "")

The mean hourly wage for the occupation in U.S. dollars for green jobs in Saint Paul is $40.65447, and for non-green jobs is $36.30688. That means green jobs in Saint Paul pay $4.347591 more than non-green jobs in Saint Paul.

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

I’d like to see a word cloud of different job titles for each sector in St. Paul

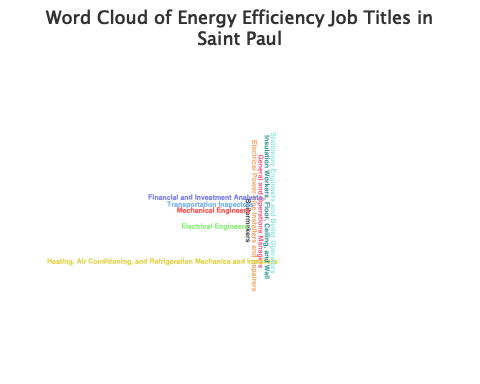
# Filter the dataset for green jobs only in St. Paul  
green\_jobs\_st\_paul <- st\_paul\_jobs %>%  
 filter(`O\*NET-SOC Sector` %in% c("Energy Efficiency", "Renewable Energy Generation", "Green Construction"))  
  
# Extract job titles and count their occurrences in St. Paul  
job\_titles\_st\_paul <- green\_jobs\_st\_paul %>%  
 count(OCC\_TITLE, sort = TRUE)  
  
# Create a word cloud for green jobs in St. Paul using highcharter  
hchart(  
 job\_titles\_st\_paul,   
 "wordcloud",   
 hcaes(name = OCC\_TITLE, weight = n)  
) %>%  
 hc\_title(text = "Word Cloud of Green Job Titles in Saint Paul")



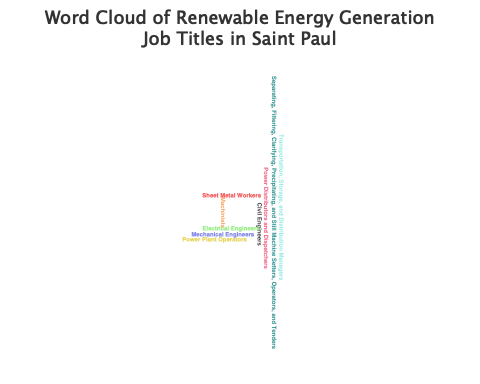
Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

Now let’s create separate word clouds for each of the green sectors (“Energy Efficiency”, “Renewable Energy Generation”, and “Green Construction”) for St. Paul

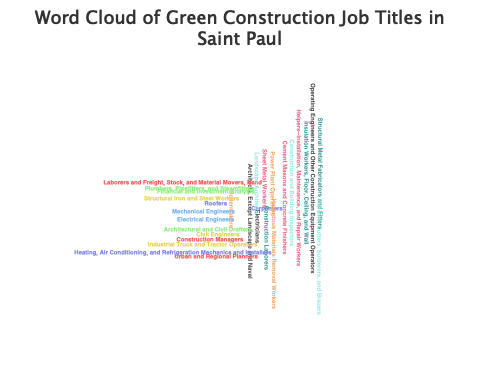
# Filter the dataset for each sector in St. Paul  
energy\_efficiency\_jobs\_st\_paul <- st\_paul\_jobs %>%  
 filter(`O\*NET-SOC Sector` == "Energy Efficiency")  
  
renewable\_energy\_jobs\_st\_paul <- st\_paul\_jobs %>%  
 filter(`O\*NET-SOC Sector` == "Renewable Energy Generation")  
  
green\_construction\_jobs\_st\_paul <- st\_paul\_jobs %>%  
 filter(`O\*NET-SOC Sector` == "Green Construction")  
  
# Create a function to generate word clouds for St. Paul sectors  
generate\_wordcloud\_st\_paul <- function(data, sector\_name) {  
 job\_titles <- data %>%  
 count(OCC\_TITLE, sort = TRUE)  
   
 hchart(  
 job\_titles,   
 "wordcloud",   
 hcaes(name = OCC\_TITLE, weight = n)  
 ) %>%  
 hc\_title(text = paste("Word Cloud of", sector\_name, "Job Titles in Saint Paul"))  
}  
  
# Generate word cloud for Energy Efficiency in St. Paul  
energy\_efficiency\_wordcloud\_st\_paul <- generate\_wordcloud\_st\_paul(energy\_efficiency\_jobs\_st\_paul, "Energy Efficiency")  
  
# Generate word cloud for Renewable Energy Generation in St. Paul  
renewable\_energy\_wordcloud\_st\_paul <- generate\_wordcloud\_st\_paul(renewable\_energy\_jobs\_st\_paul, "Renewable Energy Generation")  
  
# Generate word cloud for Green Construction in St. Paul  
green\_construction\_wordcloud\_st\_paul <- generate\_wordcloud\_st\_paul(green\_construction\_jobs\_st\_paul, "Green Construction")  
  
# Display the word clouds for St. Paul  
energy\_efficiency\_wordcloud\_st\_paul



renewable\_energy\_wordcloud\_st\_paul



green\_construction\_wordcloud\_st\_paul



Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

Let’s export for graphing

# Export the St. Paul jobs data to CSV for graphing  
write.csv(st\_paul\_jobs, here("processed\_data", "st\_paul\_jobs.csv"), row.names = FALSE)  
  
# Export the job\_type\_summary dataset for St. Paul to CSV for graphing  
write.csv(job\_type\_summary\_st\_paul, here("processed\_data", "st\_paul\_job\_type\_summary.csv"), row.names = FALSE)

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

### 3. Quality, Pay, and Qualifications of Green Jobs

|  |
| --- |
| RQ 3: What is the quality of these green jobs? How much do they pay? What qualifications are needed (education and experience) nationally? |
| **Higher education** is associated with better-quality green jobs, particularly in the energy efficiency sector. Most high- and medium-quality jobs in energy efficiency require at least a Bachelor’s Degree.  **Individuals with lower education levels** are more likely to end up in green construction, especially in lower-quality jobs. Energy efficiency tends to offer a better quality of jobs across all education levels, with strong representation in both high and medium-quality segments.  **Union membership** is **associated** with **higher-quality jobs** in all three sectors, particularly in energy efficiency and renewable energy generation, where a majority of high-quality jobs are unionized. |

# Import green job quality data  
quality\_green\_jobs <- read\_csv(here("processed\_data", "Job\_Info\_Merged\_All\_Green.csv"))

Rows: 128 Columns: 29  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (14): Reported Occupation, O\*NET-SOC Code, O\*NET-SOC Title, O\*NET-SOC Ca...  
dbl (15): Renewable Energy Generation, Energy Efficiency, Green Construction...  
  
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.

saveRDS(quality\_green\_jobs, here("processed\_data", "quality\_green\_jobs.rds"))  
  
quality\_green\_jobs <- readRDS(here("processed\_data", "quality\_green\_jobs.rds"))

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

Now, let’s transform this dataframe so it’s visualization-ready.

# Rename columns to snake\_case format  
quality\_green\_jobs <- quality\_green\_jobs %>%  
 rename(  
 report\_occupation = `Reported Occupation`,  
 onet\_soc\_code = `O\*NET-SOC Code`,  
 onet\_soc\_title = `O\*NET-SOC Title`,  
 onet\_soc\_category = `O\*NET-SOC Category`,  
 onet\_soc\_sector = `O\*NET-SOC Sector`,  
 renewable\_energy\_generation = `Renewable Energy Generation`,  
 energy\_efficiency = `Energy Efficiency`,  
 green\_construction = `Green Construction`,  
 benchmark\_total = `Benchmark Total`,  
 wage = Wage,  
 forty\_hours = `40 hours`,  
 schedule = Schedule,  
 health\_ins = `Health Ins`,  
 retirement = Retirement,  
 growth = Growth,  
 unemployment = Unemployment,  
 illness\_injury = `Illness/injury`,  
 ojt = OJT,  
 union = Union,  
 autonomy\_benchmark = autonomy\_benchmark,  
 quality = quality,  
 education = education,  
 matrix\_title = `2022 National Employment Matrix title`,  
 matrix\_code = `2022 National Employment Matrix code`,  
 typical\_education\_needed = `Typical education needed for entry`,  
 work\_experience\_related = `Work experience in a related occupation`,  
 on\_the\_job\_training = `Typical on-the-job training needed to attain competency in the occupation`  
 )

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

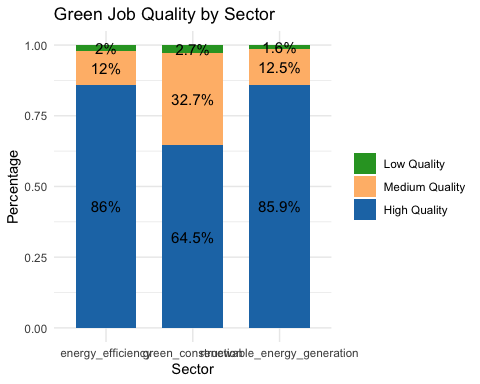
# Convert the variables to the correct types  
quality\_green\_jobs <- quality\_green\_jobs %>%  
 mutate(  
 # Factors  
 onet\_soc\_category = factor(onet\_soc\_category),  
 onet\_soc\_sector = factor(onet\_soc\_sector),  
 quality = factor(quality, levels = c("Low Quality", "Medium Quality", "High Quality")),  
 education = factor(education),  
 typical\_education\_needed = factor(typical\_education\_needed),  
 work\_experience\_related = factor(work\_experience\_related),  
 on\_the\_job\_training = factor(on\_the\_job\_training),  
   
 # Yes/No columns as numeric (1 for Yes, 0 for No)  
 renewable\_energy\_generation = as.numeric(renewable\_energy\_generation),  
 energy\_efficiency = as.numeric(energy\_efficiency),  
 green\_construction = as.numeric(green\_construction),  
 wage = as.numeric(wage),  
 forty\_hours = as.numeric(forty\_hours),  
 schedule = as.numeric(schedule),  
 health\_ins = as.numeric(health\_ins),  
 retirement = as.numeric(retirement),  
 growth = as.numeric(growth),  
 unemployment = as.numeric(unemployment),  
 illness\_injury = as.numeric(illness\_injury),  
 ojt = as.numeric(ojt),  
 union = as.numeric(union),  
 autonomy\_benchmark = as.numeric(autonomy\_benchmark)  
 )

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

# Visualize quality jobs compared across sectors  
quality\_summary <- quality\_green\_jobs %>%  
 gather(key = "sector", value = "is\_green\_job", renewable\_energy\_generation, energy\_efficiency, green\_construction) %>%  
 filter(is\_green\_job == 1) %>%  
 group\_by(sector, quality) %>%  
 summarise(count = n()) %>%  
 mutate(percentage = count / sum(count) \* 100)

`summarise()` has grouped output by 'sector'. You can override using the  
`.groups` argument.

ggplot(quality\_summary, aes(x = sector, y = percentage, fill = quality)) +  
 geom\_bar(stat = "identity", position = "fill", width = 0.7) +  
 geom\_text(aes(label = paste0(round(percentage, 1), "%")),   
 position = position\_fill(vjust = 0.5), size = 4) +  
 labs(x = "Sector", y = "Percentage", title = "Green Job Quality by Sector") +  
 scale\_fill\_manual(values = c("High Quality" = "#1f77b4", "Medium Quality" = "#ffbb78", "Low Quality" = "#2ca02c")) +  
 theme\_minimal() +  
 theme(legend.title = element\_blank())

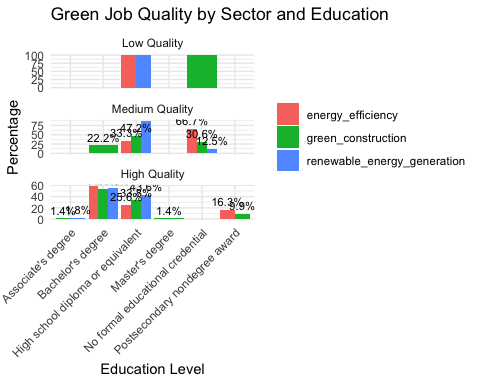


Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

# Visualize quality jobs compared for sector, education, and quality  
quality\_summary\_education <- quality\_green\_jobs %>%  
 gather(key = "sector", value = "is\_green\_job", renewable\_energy\_generation, energy\_efficiency, green\_construction) %>%  
 filter(is\_green\_job == 1) %>%  
 group\_by(sector, quality, education) %>%  
 summarise(count = n()) %>%  
 mutate(percentage = count / sum(count) \* 100)

`summarise()` has grouped output by 'sector', 'quality'. You can override using  
the `.groups` argument.

ggplot(quality\_summary\_education, aes(x = education, y = percentage, fill = sector)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 facet\_wrap(~ quality, ncol = 1, scales = "free\_y") + # Separate the quality levels  
 geom\_text(aes(label = paste0(round(percentage, 1), "%")),   
 position = position\_dodge(width = 0.9), vjust = -0.5, size = 3) +  
 labs(x = "Education Level", y = "Percentage", title = "Green Job Quality by Sector and Education") +  
 theme\_minimal() +  
 theme(legend.title = element\_blank(), axis.text.x = element\_text(angle = 45, hjust = 1))



Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

This graph shows how the distribution of green jobs (energy efficiency, green construction, renewable energy generation) varies across different levels of education, segmented by job quality (Low, Medium, High quality).

Based on the above graph:

* **Energy Efficiency** consistently dominates high-quality and medium-quality green jobs across most education levels, suggesting that this sector offers the most secure and rewarding jobs for those with varying levels of education.
* **Green Construction** has a significant presence in low-quality and medium-quality jobs, especially for those with lower education levels (such as high school diplomas or no formal educational credentials).
* **Renewable Energy Generation** seems to have fewer high-quality job opportunities compared to energy efficiency, but it does offer medium-quality opportunities, particularly for those with lower education levels.

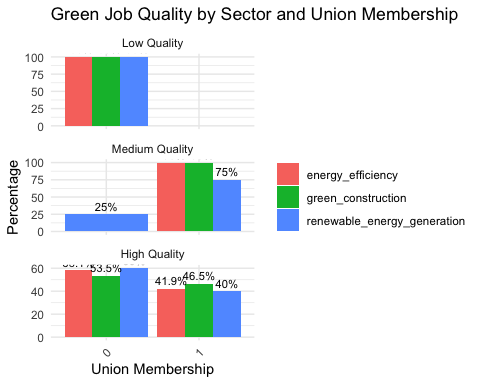
# # Matrix table that shows green job quality segmented by variables from wage to union across the three sectors (renewable\_energy\_generation, energy\_efficiency, green\_construction)  
  
# # Transform the data to long format  
# long\_quality\_jobs <- quality\_green\_jobs %>%  
# gather(key = "sector", value = "is\_green\_job", renewable\_energy\_generation, energy\_efficiency, green\_construction) %>%  
# filter(is\_green\_job == 1) %>%  
# select(sector, quality, wage, forty\_hours, schedule, health\_ins, retirement, growth, unemployment, illness\_injury, ojt, union, autonomy\_benchmark) %>%  
# pivot\_longer(cols = wage:autonomy\_benchmark, names\_to = "variable", values\_to = "value") %>%  
# group\_by(sector, quality, variable) %>%  
# summarise(proportion = mean(value)) %>%  
# ungroup() %>%  
# arrange(sector, quality, variable)  
#   
# # Generate matrix table using gt  
# matrix\_table <- long\_quality\_jobs %>%  
# pivot\_wider(names\_from = quality, values\_from = proportion) %>%  
# gt() %>%  
# tab\_header(  
# title = "Green Job Quality Matrix by Sector and Variables",  
# subtitle = "Proportion of Each Job Quality Level Across Wage, Hours, Benefits, etc."  
# ) %>%  
# fmt\_number(  
# columns = vars(`High Quality`, `Medium Quality`, `Low Quality`),  
# decimals = 2  
# )  
#   
# # Display the matrix table  
# print(matrix\_table)

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

# Visualize quality jobs compared for sector, union membership, and quality  
quality\_summary\_union <- quality\_green\_jobs %>%  
 gather(key = "sector", value = "is\_green\_job", renewable\_energy\_generation, energy\_efficiency, green\_construction) %>%  
 filter(is\_green\_job == 1) %>%  
 group\_by(sector, quality, union) %>%  
 summarise(count = n()) %>%  
 mutate(percentage = count / sum(count) \* 100)

`summarise()` has grouped output by 'sector', 'quality'. You can override using  
the `.groups` argument.

ggplot(quality\_summary\_union, aes(x = as.factor(union), y = percentage, fill = sector)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 facet\_wrap(~ quality, ncol = 1, scales = "free\_y") + # Separate the quality levels  
 geom\_text(aes(label = paste0(round(percentage, 1), "%")),   
 position = position\_dodge(width = 0.9), vjust = -0.5, size = 3) +  
 labs(x = "Union Membership", y = "Percentage", title = "Green Job Quality by Sector and Union Membership") +  
 theme\_minimal() +  
 theme(legend.title = element\_blank(), axis.text.x = element\_text(angle = 45, hjust = 1))



Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

The above graph shows green job quality segmented by union membership across three sectors: **Energy Efficiency**, **Green Construction**, and **Renewable Energy Generation**.

* **Union membership** is **associated** with **higher-quality jobs** in all three sectors, particularly in energy efficiency and renewable energy generation, where a majority of high-quality jobs are unionized. This suggests that **unions play a significant role in securing better working conditions and benefits** for workers in green jobs.
* **Medium-quality jobs** also show a clear advantage for unionized workers across the sectors, particularly in green construction and renewable energy generation.
* The presence of **union coverage even in low-quality jobs** across sectors might indicate that unionized jobs are spread across different quality categories, though the majority of benefits seem to concentrate in medium- and high-quality roles.

# Export the green quality jobs data to CSV for graphing  
saveRDS(quality\_green\_jobs, here("processed\_data", "quality\_green\_jobs.rds"))  
  
write\_csv(quality\_green\_jobs, here("processed\_data", "quality\_green\_jobs.csv"))

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

### 4. Demographics of Green Job Recipients

|  |
| --- |
| RQ 4: Who is getting these green jobs, based on education, race/ethnicity, gender, and income levels in the City of Saint Paul? |
| Of the more than **303,820 people** who live in **St. Paul**, **50.5%** are women, which is aligned with the national average. The majority of residents in St. Paul are **white** (54.3%). **Black or African American people** (15.6%) make up the largest community of color in the city. Other **communities of color,** including Asian (18.4%), Alaska and Native American (0.7%), Hispanic or Latino (8.6%) and Two or More Races (7.8%), make up about 41.5% of the population. Around **42.8% of people aged 25 and older have a bachelor’s degree** in St. Paul, which is higher than the national rate of 34% but lower than Minneapolis’ 54%. |

In the following figures, we describe the challenges women and people of color in Saint Paul are likely to face in equitably accessing the jobs that may be created through BIL and IRA funding in three specific sectors: energy efficiency, renewable-energy generation, and green construction.

* **2023 Occupational Employment and Wage Survey (OEWS):**
  + Provides employment and wage data by occupation.
  + It’s organized at the **Metropolitan Statistical Area (MSA)** level, which includes a broader geographic area (e.g., Minneapolis-St. Paul-Bloomington, MN-WI Metro).
  + This data gives insights into **jobs and wages** but lacks detailed individual demographics such as race, ethnicity, education, etc.
* **Geocorr Data from the Missouri Census Data Center:**
  + This data helps map geographic boundaries like PUMAs (Public Use Microdata Areas) to more specific local areas, such as St. Paul.
  + By using geographic weighting from Geocorr, we can estimate **St. Paul-specific** statistics from the broader MSA-level data in the OEWS.

1. Load and explore the two datasets.
2. Filter the OEWS data to the Minneapolis-St. Paul-Bloomington Metro area.
3. Get weights from the Geocorr file to adjust for St. Paul’s population.
4. Merge demographic data (from ACS) with the estimated job/wage data from OEWS.
5. Analyze the final dataset for insights into who holds green jobs in St. Paul.

**2023 Occupational Employment and Wage Survey**

✅[National level data](https://docs.google.com/spreadsheets/d/1I2munGunOJgdI2iWRW7p0BVU0O13r4zb/edit?gid=1944656488#gid=1944656488)

✅[Minneapolis-St. Paul-Bloomington, MN-WI](https://docs.google.com/spreadsheets/d/105RYiRn-1LIVC-iUdCD3fCKROfOGH_M62s45gbo7GFM/edit?gid=2141627594#gid=2141627594)

**Geocorr from the Missouri Census Data Center**

[Minnesota Level](https://docs.google.com/spreadsheets/d/1wRr-jATTjaXpErUfSCAZsoSn-lt1-TZtWv_Hoq2C29Q/edit?gid=1877088889#gid=1877088889): Contains data mapping PUMAs to Metropolitan Statistical Areas and PUMAs to cities

**Processed**

[St. Paul - ACS PUMS - Five Years + O\*NET Green Jobs](https://drive.google.com/file/d/1uyRoXlExkExjlJytEh8meqD3D--4Hwj4/view?usp=drive_link)

[National - ACS PUMS - Five Years + O\*NET Green Jobs](https://drive.google.com/file/d/1C3opSLifg144MIYnISHmf7wwZuAe2TQJ/view?usp=drive_link)

We will first need to load the **OEWS** data (Occupational Employment and Wage Survey) and the **Geocorr** data (geographic weights from the Missouri Census Data Center) into R.

* **OEWS:** processed\_data/OWES\_and\_ONET\_St\_Paul
* **Geocorr:** raw\_data/Geocorr from the Missouri Census Data Center - Minnesota.xlsx

We have the **ACS (American Community Survey)** data that provides demographic information (education, race, gender, income), so we’ll load that as well.

* **ACS:** processed\_data/St\_Paul\_ACS\_All\_Jobs.csv

**Load the OEWS data** (Occupational Employment and Wage Survey), the Geocorr data (geographic weights from the Missouri Census Data Center), and the ACS (American Community Survey) data that provides demographic information (education, race, gender, income). The OEWS Data: is already filtered to the Minneapolis-St. Paul-Bloomington, MN-WI Metro Area.

# Load St. Paul jobs data (OEWS dataset)  
st\_paul\_jobs <- read\_csv(here("processed\_data", "OWES\_and\_ONET-St\_Paul.csv"))

Rows: 742 Columns: 34  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (26): AREA\_TITLE, PRIM\_STATE, NAICS\_TITLE, I\_GROUP, OCC\_CODE, OCC\_TITLE,...  
dbl (4): AREA, AREA\_TYPE, NAICS, OWN\_CODE  
lgl (4): PCT\_TOTAL, PCT\_RPT, ANNUAL, HOURLY  
  
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.

saveRDS(st\_paul\_jobs, here("processed\_data", "st\_paul\_jobs.rds"))  
st\_paul\_jobs <- readRDS(here("processed\_data", "st\_paul\_jobs.rds"))  
  
# Load the Geocorr data from Excel (Minnesota-specific)  
geocorr\_data <- read\_excel(here("raw\_data", "Geocorr from the Missouri Census Data Center - Minnesota.xlsx"))  
  
# Load the ACS data for St. Paul (demographic data)  
acs\_data <- read\_csv(here("processed\_data", "St\_Paul\_ACS\_All\_Jobs.csv"))

Rows: 1730 Columns: 19  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (7): RT, SERIALNO, SOCP, RAC1P, SEX, SCHL, O\*NET-SOC Title  
dbl (12): DIVISION, SPORDER, PUMA20, REGION, ST, AGEP, PINCP, ADJINC, WAGP, ...  
  
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.

saveRDS(acs\_data, here("processed\_data", "acs\_data.rds"))  
acs\_data <- readRDS(here("processed\_data", "acs\_data.rds"))

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

**Use Geocorr to apply weights.** The **Geocorr** data provides the geographic weights that represent the population of St. Paul relative to the larger metro area.The purpose of this step is to adjust the OEWS data to better represent **St. Paul** specifically. We will calculate the percentage of St. Paul’s population in the larger metro area using **Geocorr** and apply this as a weight to the OEWS data.

Since we are working with PUMAs (Public Use Microdata Areas) and need to adjust the OEWS data for St. Paul using the Geocorr weights, we’ll focus on using:

* Total population (2020 Census) to understand the population of each PUMA.
* puma22-to-cbsa20 allocation factor, which represents the proportion of the population in each PUMA that falls within the Minneapolis-St. Paul-Bloomington metro area.

The St. Paul weighted allocation factor is 1, which means that for these specific PUMAs (representing **Ramsey County–St. Paul City**), the entire population is considered part of the Minneapolis-St. Paul-Bloomington metro area. This result suggests that we can apply the full population of these PUMAs without needing further weighting, which simplifies the next steps.

# Filter the Geocorr data to only include St. Paul PUMAs  
st\_paul\_geocorr <- geocorr\_data %>%  
 filter(`PUMA22 name` %in% c(  
 "Ramsey County--St. Paul City (Northwest)",   
 "Ramsey County--St. Paul City (Southwest)",   
 "Ramsey County--St. Paul City (East)"  
 ))  
  
# Calculate the total population for St. Paul PUMAs and the weighted allocation factor  
st\_paul\_population <- sum(st\_paul\_geocorr$`Total population (2020 Census)`, na.rm = TRUE)  
metro\_population <- sum(geocorr\_data$`Total population (2020 Census)`, na.rm = TRUE)  
  
# Calculate the weighted population factor for St. Paul within the metro area using the allocation factor  
st\_paul\_weight <- sum(st\_paul\_geocorr$`puma22-to-cbsa20 allocation factor`, na.rm = TRUE) / nrow(st\_paul\_geocorr)  
  
# Output the results  
cat("St. Paul weighted allocation factor:", st\_paul\_weight, "\n")

St. Paul weighted allocation factor: 1

cat("St. Paul population:", format(st\_paul\_population, big.mark = ",", scientific = FALSE), "\n")

St. Paul population: 311,527

cat("Metro area population:", format(metro\_population, big.mark = ",", scientific = FALSE), "\n")

Metro area population: 5,706,494

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

**Estimate St. Paul-specific data**. Multiply the employment numbers and wages in the **OEWS** data by the St. Paul weight to get **St. Paul-specific** employment and wage estimates.

We’ll multiply the employment numbers (TOT\_EMP) and wages (H\_MEAN and A\_MEAN) from the **OEWS** dataset by this St. Paul weight to get **St. Paul-specific estimates**.

# Ensure that the necessary columns are numeric  
st\_paul\_jobs <- st\_paul\_jobs %>%  
 mutate(  
 TOT\_EMP = as.numeric(TOT\_EMP), # Convert total employment to numeric  
 H\_MEAN = as.numeric(H\_MEAN), # Convert mean hourly wage to numeric  
 A\_MEAN = as.numeric(A\_MEAN) # Convert mean annual wage to numeric  
 )

Warning: There were 3 warnings in `mutate()`.  
The first warning was:  
ℹ In argument: `TOT\_EMP = as.numeric(TOT\_EMP)`.  
Caused by warning:  
! NAs introduced by coercion  
ℹ Run `dplyr::last\_dplyr\_warnings()` to see the 2 remaining warnings.

# Apply the St. Paul weight to the OEWS dataset to adjust the employment and wage data for St. Paul  
st\_paul\_jobs\_weighted <- st\_paul\_jobs %>%  
 mutate(  
 TOT\_EMP\_St\_Paul = TOT\_EMP \* st\_paul\_weight, # Adjusting total employment to St. Paul  
 H\_MEAN\_St\_Paul = H\_MEAN \* st\_paul\_weight, # Adjusting mean hourly wage  
 A\_MEAN\_St\_Paul = A\_MEAN \* st\_paul\_weight # Adjusting mean annual wage  
 )  
  
# Output a glimpse of the adjusted St. Paul-specific job data  
glimpse(st\_paul\_jobs\_weighted)

Rows: 742  
Columns: 37  
$ AREA <dbl> 33460, 33460, 33460, 33460, 33460, 33460, 33460, 33…  
$ AREA\_TITLE <chr> "Minneapolis-St. Paul-Bloomington, MN-WI", "Minneap…  
$ AREA\_TYPE <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, …  
$ PRIM\_STATE <chr> "MN", "MN", "MN", "MN", "MN", "MN", "MN", "MN", "MN…  
$ NAICS <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
$ NAICS\_TITLE <chr> "Cross-industry", "Cross-industry", "Cross-industry…  
$ I\_GROUP <chr> "cross-industry", "cross-industry", "cross-industry…  
$ OWN\_CODE <dbl> 1235, 1235, 1235, 1235, 1235, 1235, 1235, 1235, 123…  
$ OCC\_CODE <chr> "00-0000", "11-0000", "11-1011", "11-1021", "11-103…  
$ OCC\_TITLE <chr> "All Occupations", "Management Occupations", "Chief…  
$ O\_GROUP <chr> "total", "major", "detailed", "detailed", "detailed…  
$ TOT\_EMP <dbl> 1911030, 140870, 4420, 48300, 70, 90, 7000, 7390, 9…  
$ EMP\_PRSE <chr> "0", "1.2", "3.9", "2", "11.9", "20.5", "3.4", "11.…  
$ JOBS\_1000 <chr> "1000", "73.712", "2.315", "25.272", "0.036", "0.04…  
$ LOC\_QUOTIENT <chr> "1", "1.07", "1.66", "1.09", "0.17", "0.36", "1.51"…  
$ PCT\_TOTAL <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…  
$ PCT\_RPT <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…  
$ H\_MEAN <dbl> 33.80, 69.10, 129.79, 59.93, NA, 59.25, 86.25, 81.1…  
$ A\_MEAN <dbl> 70290, 143730, 269950, 124650, 82650, 123240, 17939…  
$ MEAN\_PRSE <chr> "0.5", "1", "3.5", "1.2", "2.6", "8.7", "1.9", "2.4…  
$ H\_PCT10 <chr> "15.06", "29.27", "45.92", "22.92", "\*", "33.65", "…  
$ H\_PCT25 <chr> "18.45", "41.06", "67.35", "32.21", "\*", "37.64", "…  
$ H\_MEDIAN <chr> "26.37", "61.33", "100.87", "49.26", "\*", "55.5", "…  
$ H\_PCT75 <chr> "39.85", "83.45", "#", "76.63", "\*", "70.61", "101.…  
$ H\_PCT90 <chr> "60.7", "110.83", "#", "105.67", "\*", "98.35", "#",…  
$ A\_PCT10 <chr> "31320", "60890", "95510", "47670", "21140", "69990…  
$ A\_PCT25 <chr> "38380", "85410", "140090", "66990", "37050", "7828…  
$ A\_MEDIAN <chr> "54850", "127570", "209820", "102460", "79000", "11…  
$ A\_PCT75 <chr> "82890", "173570", "#", "159380", "114000", "146860…  
$ A\_PCT90 <chr> "126260", "230520", "#", "219800", "151910", "20457…  
$ ANNUAL <lgl> NA, NA, NA, NA, TRUE, NA, NA, NA, NA, NA, NA, NA, N…  
$ HOURLY <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…  
$ `O\*NET-SOC Code` <chr> NA, NA, NA, "11-1021", NA, NA, NA, NA, NA, NA, NA, …  
$ `O\*NET-SOC Sector` <chr> NA, NA, NA, "Energy Efficiency", NA, NA, NA, NA, NA…  
$ TOT\_EMP\_St\_Paul <dbl> 1911030, 140870, 4420, 48300, 70, 90, 7000, 7390, 9…  
$ H\_MEAN\_St\_Paul <dbl> 33.80, 69.10, 129.79, 59.93, NA, 59.25, 86.25, 81.1…  
$ A\_MEAN\_St\_Paul <dbl> 70290, 143730, 269950, 124650, 82650, 123240, 17939…

# Save the adjusted St. Paul-specific dataset to an RDS and CSV file for future analysis  
saveRDS(st\_paul\_jobs\_weighted, here("processed\_data", "st\_paul\_jobs\_weighted.rds"))  
write\_csv(st\_paul\_jobs\_weighted, here("processed\_data", "st\_paul\_jobs\_weighted.csv"))  
  
# Optional: Print some summary statistics for St. Paul-specific employment and wages  
summary\_st\_paul\_jobs <- st\_paul\_jobs\_weighted %>%  
 summarize(  
 Total\_Employment = sum(TOT\_EMP\_St\_Paul, na.rm = TRUE),  
 Mean\_Hourly\_Wage = mean(H\_MEAN\_St\_Paul, na.rm = TRUE),  
 Mean\_Annual\_Wage = mean(A\_MEAN\_St\_Paul, na.rm = TRUE)  
 )  
  
# Output the summary for quick analysis  
print(summary\_st\_paul\_jobs)

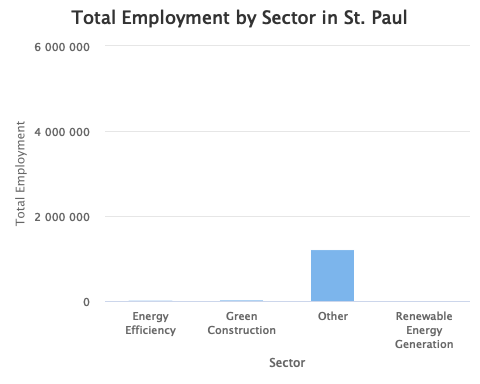
# A tibble: 1 × 3  
 Total\_Employment Mean\_Hourly\_Wage Mean\_Annual\_Wage  
 <dbl> <dbl> <dbl>  
1 5751890 36.6 77675.

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

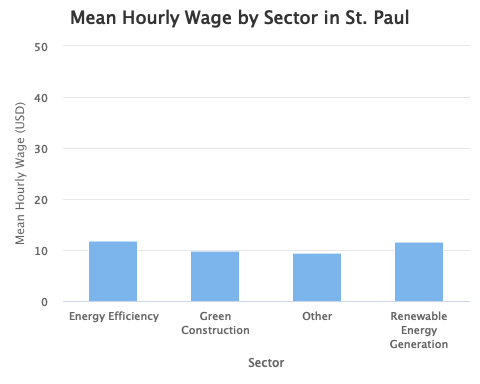
# Replace NA values in 'O\*NET-SOC Sector' with 'Other'  
st\_paul\_jobs\_weighted <- st\_paul\_jobs\_weighted %>%  
 mutate(`O\*NET-SOC Sector` = ifelse(is.na(`O\*NET-SOC Sector`), "Other", `O\*NET-SOC Sector`))  
  
# Group by 'O\*NET-SOC Sector' and calculate the total employment, mean hourly wage, and mean annual wage  
sector\_summary\_st\_paul <- st\_paul\_jobs\_weighted %>%  
 group\_by(`O\*NET-SOC Sector`) %>%  
 summarize(  
 Total\_Employment = sum(TOT\_EMP\_St\_Paul, na.rm = TRUE),  
 Mean\_Hourly\_Wage = mean(H\_MEAN\_St\_Paul, na.rm = TRUE),  
 Mean\_Annual\_Wage = mean(A\_MEAN\_St\_Paul, na.rm = TRUE)  
 )  
  
# Output the sector-wise summary  
print(sector\_summary\_st\_paul)

# A tibble: 4 × 4  
 `O\*NET-SOC Sector` Total\_Employment Mean\_Hourly\_Wage Mean\_Annual\_Wage  
 <chr> <dbl> <dbl> <dbl>  
1 Energy Efficiency 66410 45.5 94669.  
2 Green Construction 124680 37.9 78809.  
3 Other 5537550 36.3 77193.  
4 Renewable Energy Generation 23250 44.7 92991.

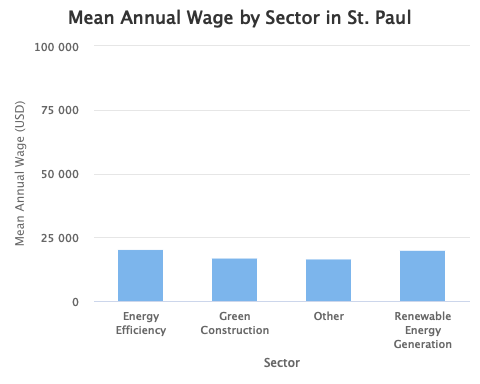
# Save the sector-wise summary as an RDS and CSV file for future reference  
saveRDS(sector\_summary\_st\_paul, here("processed\_data", "sector\_summary\_st\_paul.rds"))  
write\_csv(sector\_summary\_st\_paul, here("processed\_data", "sector\_summary\_st\_paul.csv"))  
  
# Visualization for Total Employment across sectors  
hchart(sector\_summary\_st\_paul, "column", hcaes(x = `O\*NET-SOC Sector`, y = Total\_Employment)) %>%  
 hc\_title(text = "Total Employment by Sector in St. Paul") %>%  
 hc\_xAxis(title = list(text = "Sector")) %>%  
 hc\_yAxis(title = list(text = "Total Employment"), labels = list(format = "{value:,0f}")) %>%  
 hc\_tooltip(pointFormat = '<b>{point.y:,0f}</b>')



# Visualization for Mean Hourly Wage across sectors  
hchart(sector\_summary\_st\_paul, "column", hcaes(x = `O\*NET-SOC Sector`, y = Mean\_Hourly\_Wage)) %>%  
 hc\_title(text = "Mean Hourly Wage by Sector in St. Paul") %>%  
 hc\_xAxis(title = list(text = "Sector")) %>%  
 hc\_yAxis(title = list(text = "Mean Hourly Wage (USD)")) %>%  
 hc\_tooltip(pointFormat = '<b>{point.y:.2f} USD</b>')



# Visualization for Mean Annual Wage across sectors  
hchart(sector\_summary\_st\_paul, "column", hcaes(x = `O\*NET-SOC Sector`, y = Mean\_Annual\_Wage)) %>%  
 hc\_title(text = "Mean Annual Wage by Sector in St. Paul") %>%  
 hc\_xAxis(title = list(text = "Sector")) %>%  
 hc\_yAxis(title = list(text = "Mean Annual Wage (USD)"), labels = list(format = "{value:,0f}")) %>%  
 hc\_tooltip(pointFormat = '<b>{point.y:,0f} USD</b>')



Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

**Incorporate ACS demographics**. We will merge the OEWS data with the **ACS** data. The **ACS data** has demographic information like education, race/ethnicity, gender, and income levels. This will allow us to analyze the green job data segmented by these demographic factors in **St. Paul**.

# Convert the O\*NET-SOC code to character in both datasets  
st\_paul\_jobs\_weighted <- st\_paul\_jobs\_weighted %>%  
 mutate(`O\*NET-SOC Code` = as.character(`O\*NET-SOC Code`))  
  
acs\_data <- acs\_data %>%  
 mutate(`O\*NET-SOC Code` = as.character(`O\*NET-SOC Code`))

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

# Filter the ACS data to only include rows where 'Green Job Flag' is 1  
acs\_green\_data <- acs\_data %>%   
 filter(`Green Job Flag` == 1)  
  
# Check the filtered data to ensure it looks correct  
glimpse(acs\_green\_data)

Rows: 72  
Columns: 19  
$ RT <chr> "P", "P", "P", "P", "P", "P", "P", "P", "P", "P", "P…  
$ SERIALNO <chr> "2022GQ0001538", "2022GQ0013624", "2022GQ0025479", "…  
$ DIVISION <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4…  
$ SPORDER <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 5, 1, 1, 2, 2, 2…  
$ PUMA20 <dbl> 1505, 1504, 1504, 1503, 1503, 1505, 1504, 1504, 1504…  
$ REGION <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2…  
$ ST <dbl> 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, …  
$ AGEP <dbl> 54, 19, 18, 50, 20, 42, 18, 18, 50, 20, 21, 67, 45, …  
$ SOCP <chr> "472181", "537062", "537062", "537062", "537062", "4…  
$ RAC1P <chr> "White alone", "Asian alone", "Two or More Races", "…  
$ SEX <chr> "Male", "Male", "Male", "Male", "Male", "Male", "Mal…  
$ SCHL <chr> "GED or alternative credential", "Regular high schoo…  
$ PINCP <dbl> 0, 4000, 4000, 19200, 4000, 14700, 4000, 20000, 1920…  
$ ADJINC <dbl> 1042311, 1042311, 1042311, 1042311, 1042311, 1042311…  
$ WAGP <dbl> 0, 4000, 4000, 18000, 4000, 11100, 4000, 20000, 1800…  
$ PWGTP <dbl> 1, 5, 12, 12, 8, 1, 5, 7, 12, 12, 14, 30, 35, 33, 29…  
$ `O\*NET-SOC Code` <chr> "472181", "537062", "537062", "537062", "537062", "4…  
$ `O\*NET-SOC Title` <chr> "Roofers", "Laborers and Freight, Stock, and Materia…  
$ `Green Job Flag` <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

# Remove rows with NA O\*NET-SOC Codes in both datasets before merging  
st\_paul\_jobs\_weighted <- st\_paul\_jobs\_weighted %>%  
 filter(!is.na(`O\*NET-SOC Code`))  
  
acs\_green\_data <- acs\_green\_data %>%  
 filter(!is.na(`O\*NET-SOC Code`))  
  
# Re-check the unique O\*NET-SOC Codes after filtering out NA values  
unique\_jobs\_codes <- unique(st\_paul\_jobs\_weighted$`O\*NET-SOC Code`)  
unique\_acs\_codes <- unique(acs\_green\_data$`O\*NET-SOC Code`)  
  
# Find codes that exist in one dataset but not the other  
missing\_in\_acs <- setdiff(unique\_jobs\_codes, unique\_acs\_codes)  
missing\_in\_jobs <- setdiff(unique\_acs\_codes, unique\_jobs\_codes)  
  
# Output the results  
cat("Codes in jobs but not in ACS:", missing\_in\_acs, "\n")

Codes in jobs but not in ACS: 11-1021 11-3071 11-9021 13-2051 17-1011 17-1012 17-2051 17-2071 17-2141 17-3011 19-3051 47-2011 47-2031 47-2051 47-2061 47-2073 47-2111 47-2131 47-2152 47-2181 47-2211 47-2221 47-4011 47-4041 49-9021 49-9051 49-9098 51-2041 51-4041 51-4121 51-8012 51-8013 51-8021 51-9012 53-6051 53-7051 53-7062 47-5041 49-9042 19-4051 51-8011 19-4041 47-5013 13-1073 47-3012

cat("Codes in ACS but not in jobs:", missing\_in\_jobs, "\n")

Codes in ACS but not in jobs: 472181 537062 472061 514041 111021 472111 132051 537051 172141 172051 472152 472031 474011 113071 171011 119021 518021

# Ensure the O\*NET-SOC Code format is consistent  
# Remove hyphens from the codes in both datasets for consistent matching  
st\_paul\_jobs\_weighted <- st\_paul\_jobs\_weighted %>%  
 mutate(`O\*NET-SOC Code` = gsub("-", "", `O\*NET-SOC Code`))  
  
acs\_green\_data <- acs\_green\_data %>%  
 mutate(`O\*NET-SOC Code` = gsub("-", "", `O\*NET-SOC Code`))  
  
# Re-check the unique O\*NET-SOC Codes after formatting  
unique\_jobs\_codes <- unique(st\_paul\_jobs\_weighted$`O\*NET-SOC Code`)  
unique\_acs\_codes <- unique(acs\_green\_data$`O\*NET-SOC Code`)  
  
# Find codes that exist in one dataset but not the other  
missing\_in\_acs <- setdiff(unique\_jobs\_codes, unique\_acs\_codes)  
missing\_in\_jobs <- setdiff(unique\_acs\_codes, unique\_jobs\_codes)  
  
# Output the results  
cat("Codes in jobs but not in ACS:", missing\_in\_acs, "\n")

Codes in jobs but not in ACS: 171012 172071 173011 193051 472011 472051 472073 472131 472211 472221 474041 499021 499051 499098 512041 514121 518012 518013 519012 536051 475041 499042 194051 518011 194041 475013 131073 473012

cat("Codes in ACS but not in jobs:", missing\_in\_jobs, "\n")

Codes in ACS but not in jobs:

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

# Merge the 'st\_paul\_jobs\_weighted' data with 'acs\_green\_data' using the 'O\*NET-SOC Code' column  
merged\_green\_jobs\_data <- st\_paul\_jobs\_weighted %>%  
 left\_join(acs\_green\_data, by = "O\*NET-SOC Code")

Warning in left\_join(., acs\_green\_data, by = "O\*NET-SOC Code"): Detected an unexpected many-to-many relationship between `x` and `y`.  
ℹ Row 1 of `x` matches multiple rows in `y`.  
ℹ Row 17 of `y` matches multiple rows in `x`.  
ℹ If a many-to-many relationship is expected, set `relationship =  
 "many-to-many"` to silence this warning.

# Check the merged data to validate the merge  
glimpse(merged\_green\_jobs\_data)

Rows: 124  
Columns: 55  
$ AREA <dbl> 33460, 33460, 33460, 33460, 33460, 33460, 33460, 33…  
$ AREA\_TITLE <chr> "Minneapolis-St. Paul-Bloomington, MN-WI", "Minneap…  
$ AREA\_TYPE <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, …  
$ PRIM\_STATE <chr> "MN", "MN", "MN", "MN", "MN", "MN", "MN", "MN", "MN…  
$ NAICS <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
$ NAICS\_TITLE <chr> "Cross-industry", "Cross-industry", "Cross-industry…  
$ I\_GROUP <chr> "cross-industry", "cross-industry", "cross-industry…  
$ OWN\_CODE <dbl> 1235, 1235, 1235, 1235, 1235, 1235, 1235, 1235, 123…  
$ OCC\_CODE <chr> "11-1021", "11-1021", "11-1021", "11-1021", "11-102…  
$ OCC\_TITLE <chr> "General and Operations Managers", "General and Ope…  
$ O\_GROUP <chr> "detailed", "detailed", "detailed", "detailed", "de…  
$ TOT\_EMP <dbl> 48300, 48300, 48300, 48300, 48300, 48300, 2610, 261…  
$ EMP\_PRSE <chr> "2", "2", "2", "2", "2", "2", "3.1", "3.1", "3.1", …  
$ JOBS\_1000 <chr> "25.272", "25.272", "25.272", "25.272", "25.272", "…  
$ LOC\_QUOTIENT <chr> "1.09", "1.09", "1.09", "1.09", "1.09", "1.09", "1.…  
$ PCT\_TOTAL <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…  
$ PCT\_RPT <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…  
$ H\_MEAN <dbl> 59.93, 59.93, 59.93, 59.93, 59.93, 59.93, 61.90, 61…  
$ A\_MEAN <dbl> 124650, 124650, 124650, 124650, 124650, 124650, 128…  
$ MEAN\_PRSE <chr> "1.2", "1.2", "1.2", "1.2", "1.2", "1.2", "1.1", "1…  
$ H\_PCT10 <chr> "22.92", "22.92", "22.92", "22.92", "22.92", "22.92…  
$ H\_PCT25 <chr> "32.21", "32.21", "32.21", "32.21", "32.21", "32.21…  
$ H\_MEDIAN <chr> "49.26", "49.26", "49.26", "49.26", "49.26", "49.26…  
$ H\_PCT75 <chr> "76.63", "76.63", "76.63", "76.63", "76.63", "76.63…  
$ H\_PCT90 <chr> "105.67", "105.67", "105.67", "105.67", "105.67", "…  
$ A\_PCT10 <chr> "47670", "47670", "47670", "47670", "47670", "47670…  
$ A\_PCT25 <chr> "66990", "66990", "66990", "66990", "66990", "66990…  
$ A\_MEDIAN <chr> "102460", "102460", "102460", "102460", "102460", "…  
$ A\_PCT75 <chr> "159380", "159380", "159380", "159380", "159380", "…  
$ A\_PCT90 <chr> "219800", "219800", "219800", "219800", "219800", "…  
$ ANNUAL <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…  
$ HOURLY <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…  
$ `O\*NET-SOC Code` <chr> "111021", "111021", "111021", "111021", "111021", "…  
$ `O\*NET-SOC Sector` <chr> "Energy Efficiency", "Energy Efficiency", "Energy E…  
$ TOT\_EMP\_St\_Paul <dbl> 48300, 48300, 48300, 48300, 48300, 48300, 2610, 261…  
$ H\_MEAN\_St\_Paul <dbl> 59.93, 59.93, 59.93, 59.93, 59.93, 59.93, 61.90, 61…  
$ A\_MEAN\_St\_Paul <dbl> 124650, 124650, 124650, 124650, 124650, 124650, 128…  
$ RT <chr> "P", "P", "P", "P", "P", "P", "P", "P", "P", "P", "…  
$ SERIALNO <chr> "2022HU0005354", "2022HU0054042", "2022HU0097062", …  
$ DIVISION <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, …  
$ SPORDER <dbl> 2, 1, 1, 2, 1, 2, 2, 1, 1, 1, 3, 2, 1, 1, 2, 1, 1, …  
$ PUMA20 <dbl> 1504, 1504, 1505, 1504, 1505, 1504, 1505, 1505, 150…  
$ REGION <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, …  
$ ST <dbl> 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,…  
$ AGEP <dbl> 67, 35, 30, 46, 64, 37, 51, 39, 22, 34, 26, 31, 60,…  
$ SOCP <chr> "111021", "111021", "111021", "111021", "111021", "…  
$ RAC1P <chr> "White alone", "White alone", "Black or African Ame…  
$ SEX <chr> "Male", "Female", "Male", "Female", "Female", "Male…  
$ SCHL <chr> "1 or more years of college credit, no degree", "Ba…  
$ PINCP <dbl> 41000, 50000, 38000, 30000, 0, 94000, 68000, 120000…  
$ ADJINC <dbl> 1042311, 1042311, 1042311, 1042311, 1042311, 104231…  
$ WAGP <dbl> 14000, 50000, 38000, 0, 0, 94000, 68000, 120000, 20…  
$ PWGTP <dbl> 30, 29, 175, 25, 12, 25, 18, 35, 14, 28, 76, 27, 32…  
$ `O\*NET-SOC Title` <chr> "General and Operations Managers", "General and Ope…  
$ `Green Job Flag` <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …

# Save the merged dataset for future reference  
saveRDS(merged\_green\_jobs\_data, here("processed\_data", "merged\_green\_jobs\_data.rds"))  
write\_csv(merged\_green\_jobs\_data, here("processed\_data", "merged\_green\_jobs\_data.csv"))

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

# Assess the number of duplicates in the merged dataset by counting occurrences of each O\*NET-SOC Code  
duplication\_summary <- merged\_green\_jobs\_data %>%  
 group\_by(`O\*NET-SOC Code`) %>%  
 summarise(count = n()) %>%  
 arrange(desc(count))  
  
# View the summary of duplication  
print(duplication\_summary)

# A tibble: 45 × 2  
 `O\*NET-SOC Code` count  
 <chr> <int>  
 1 537062 21  
 2 172141 12  
 3 472061 8  
 4 472111 7  
 5 111021 6  
 6 132051 6  
 7 172051 6  
 8 113071 3  
 9 171011 3  
10 172071 3  
# ℹ 35 more rows

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

If we need accurate totals or averages across jobs and demographics, **aggregation** will be necessary to avoid inflating the data.

If we feel the current level of detail (with the duplicates) provides useful insights, we can keep the data as is but be mindful of how you interpret summed metrics. This is what we will do. The duplication is meaningful (for example, because a job can truly exist in multiple sectors or demographics are validly associated with multiple jobs), we choose to keep the dataset as is. This would allow us to analyze the data with all the overlaps. However, we need to be cautious that this doesn't skew metrics that sum values (like total employment).

# # Aggregate demographic and job-related data by O\*NET-SOC Code and O\*NET-SOC Sector after merge  
# aggregated\_data <- merged\_green\_jobs\_data %>%  
# group\_by(`O\*NET-SOC Code`, `O\*NET-SOC Sector`, `O\*NET-SOC Title`) %>%  
# summarise(  
# Total\_Employment = sum(TOT\_EMP\_St\_Paul, na.rm = TRUE),  
# Mean\_Hourly\_Wage = mean(H\_MEAN\_St\_Paul, na.rm = TRUE),  
# Mean\_Annual\_Wage = mean(A\_MEAN\_St\_Paul, na.rm = TRUE),  
# Average\_Age = mean(AGEP, na.rm = TRUE),  
# Proportion\_Female = mean(SEX == "Female", na.rm = TRUE),  
# Proportion\_Male = mean(SEX == "Male", na.rm = TRUE),  
# Average\_Income = mean(PINCP, na.rm = TRUE),  
# Proportion\_White = mean(RAC1P == "White alone", na.rm = TRUE),  
# Proportion\_Black = mean(RAC1P == "Black or African American alone", na.rm = TRUE),  
# Count = n() # This column helps you see how many records were combined in this aggregation  
# )  
#   
# # View the aggregated data  
# glimpse(aggregated\_data)

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

**Analyze the data**. Once the datasets are merged, you can start analyzing the data to answer your research question.

Now that we’ve merged the **OEWS** and **ACS** data, we can group by the **O\*NET-SOC Sector** (Energy Efficiency, Renewable Energy Generation, Green Construction) and demographic factors like **education**, **race**, **gender**, and **income**.

# Convert data types  
merged\_green\_jobs\_data <- merged\_green\_jobs\_data %>%  
 mutate(  
 NAICS\_TITLE = as.factor(NAICS\_TITLE), # Factor for industry titles  
 I\_GROUP = as.factor(I\_GROUP), # Factor for industry group  
 O\_GROUP = as.factor(O\_GROUP), # Factor for occupation group  
 H\_PCT10 = as.numeric(H\_PCT10), # Convert percentages to numeric  
 H\_PCT25 = as.numeric(H\_PCT25),  
 H\_MEDIAN = as.numeric(H\_MEDIAN),  
 H\_PCT75 = as.numeric(H\_PCT75),  
 H\_PCT90 = as.numeric(H\_PCT90),  
 A\_PCT10 = as.numeric(A\_PCT10),  
 A\_PCT25 = as.numeric(A\_PCT25),  
 A\_MEDIAN = as.numeric(A\_MEDIAN),  
 A\_PCT75 = as.numeric(A\_PCT75),  
 A\_PCT90 = as.numeric(A\_PCT90),  
 ANNUAL = as.numeric(ANNUAL), # Convert to numeric for consistency  
 HOURLY = as.numeric(HOURLY),  
 `O\*NET-SOC Sector` = as.factor(`O\*NET-SOC Sector`), # Factor for green job sectors  
 TOT\_EMP\_St\_Paul = as.numeric(TOT\_EMP\_St\_Paul), # Numeric for employment totals  
 H\_MEAN\_St\_Paul = as.numeric(H\_MEAN\_St\_Paul), # Numeric for hourly wage in St. Paul  
 A\_MEAN\_St\_Paul = as.numeric(A\_MEAN\_St\_Paul), # Numeric for annual wage in St. Paul  
 AGEP = as.factor(AGEP), # Age as a factor if we treat it categorically  
 RAC1P = as.factor(RAC1P), # Factor for race  
 SEX = as.factor(SEX), # Factor for gender  
 SCHL = as.factor(SCHL), # Factor for education level  
 PINCP = as.numeric(PINCP), # Numeric for personal income  
 ADJINC = as.numeric(ADJINC), # Numeric for adjusted income  
 WAGP = as.numeric(WAGP), # Numeric for wage  
 PWGTP = as.numeric(PWGTP), # Numeric for person weight  
 `Green Job Flag` = as.numeric(`Green Job Flag`) # Yes/No as numeric  
 )

Warning: There were 10 warnings in `mutate()`.  
The first warning was:  
ℹ In argument: `H\_PCT10 = as.numeric(H\_PCT10)`.  
Caused by warning:  
! NAs introduced by coercion  
ℹ Run `dplyr::last\_dplyr\_warnings()` to see the 9 remaining warnings.

# Double-check the changes  
glimpse(merged\_green\_jobs\_data)

Rows: 124  
Columns: 55  
$ AREA <dbl> 33460, 33460, 33460, 33460, 33460, 33460, 33460, 33…  
$ AREA\_TITLE <chr> "Minneapolis-St. Paul-Bloomington, MN-WI", "Minneap…  
$ AREA\_TYPE <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, …  
$ PRIM\_STATE <chr> "MN", "MN", "MN", "MN", "MN", "MN", "MN", "MN", "MN…  
$ NAICS <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
$ NAICS\_TITLE <fct> Cross-industry, Cross-industry, Cross-industry, Cro…  
$ I\_GROUP <fct> cross-industry, cross-industry, cross-industry, cro…  
$ OWN\_CODE <dbl> 1235, 1235, 1235, 1235, 1235, 1235, 1235, 1235, 123…  
$ OCC\_CODE <chr> "11-1021", "11-1021", "11-1021", "11-1021", "11-102…  
$ OCC\_TITLE <chr> "General and Operations Managers", "General and Ope…  
$ O\_GROUP <fct> detailed, detailed, detailed, detailed, detailed, d…  
$ TOT\_EMP <dbl> 48300, 48300, 48300, 48300, 48300, 48300, 2610, 261…  
$ EMP\_PRSE <chr> "2", "2", "2", "2", "2", "2", "3.1", "3.1", "3.1", …  
$ JOBS\_1000 <chr> "25.272", "25.272", "25.272", "25.272", "25.272", "…  
$ LOC\_QUOTIENT <chr> "1.09", "1.09", "1.09", "1.09", "1.09", "1.09", "1.…  
$ PCT\_TOTAL <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…  
$ PCT\_RPT <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…  
$ H\_MEAN <dbl> 59.93, 59.93, 59.93, 59.93, 59.93, 59.93, 61.90, 61…  
$ A\_MEAN <dbl> 124650, 124650, 124650, 124650, 124650, 124650, 128…  
$ MEAN\_PRSE <chr> "1.2", "1.2", "1.2", "1.2", "1.2", "1.2", "1.1", "1…  
$ H\_PCT10 <dbl> 22.92, 22.92, 22.92, 22.92, 22.92, 22.92, 32.84, 32…  
$ H\_PCT25 <dbl> 32.21, 32.21, 32.21, 32.21, 32.21, 32.21, 41.03, 41…  
$ H\_MEDIAN <dbl> 49.26, 49.26, 49.26, 49.26, 49.26, 49.26, 53.89, 53…  
$ H\_PCT75 <dbl> 76.63, 76.63, 76.63, 76.63, 76.63, 76.63, 74.34, 74…  
$ H\_PCT90 <dbl> 105.67, 105.67, 105.67, 105.67, 105.67, 105.67, 99.…  
$ A\_PCT10 <dbl> 47670, 47670, 47670, 47670, 47670, 47670, 68310, 68…  
$ A\_PCT25 <dbl> 66990, 66990, 66990, 66990, 66990, 66990, 85340, 85…  
$ A\_MEDIAN <dbl> 102460, 102460, 102460, 102460, 102460, 102460, 112…  
$ A\_PCT75 <dbl> 159380, 159380, 159380, 159380, 159380, 159380, 154…  
$ A\_PCT90 <dbl> 219800, 219800, 219800, 219800, 219800, 219800, 207…  
$ ANNUAL <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…  
$ HOURLY <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…  
$ `O\*NET-SOC Code` <chr> "111021", "111021", "111021", "111021", "111021", "…  
$ `O\*NET-SOC Sector` <fct> Energy Efficiency, Energy Efficiency, Energy Effici…  
$ TOT\_EMP\_St\_Paul <dbl> 48300, 48300, 48300, 48300, 48300, 48300, 2610, 261…  
$ H\_MEAN\_St\_Paul <dbl> 59.93, 59.93, 59.93, 59.93, 59.93, 59.93, 61.90, 61…  
$ A\_MEAN\_St\_Paul <dbl> 124650, 124650, 124650, 124650, 124650, 124650, 128…  
$ RT <chr> "P", "P", "P", "P", "P", "P", "P", "P", "P", "P", "…  
$ SERIALNO <chr> "2022HU0005354", "2022HU0054042", "2022HU0097062", …  
$ DIVISION <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, …  
$ SPORDER <dbl> 2, 1, 1, 2, 1, 2, 2, 1, 1, 1, 3, 2, 1, 1, 2, 1, 1, …  
$ PUMA20 <dbl> 1504, 1504, 1505, 1504, 1505, 1504, 1505, 1505, 150…  
$ REGION <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, …  
$ ST <dbl> 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,…  
$ AGEP <fct> 67, 35, 30, 46, 64, 37, 51, 39, 22, 34, 26, 31, 60,…  
$ SOCP <chr> "111021", "111021", "111021", "111021", "111021", "…  
$ RAC1P <fct> White alone, White alone, Black or African American…  
$ SEX <fct> Male, Female, Male, Female, Female, Male, Male, Mal…  
$ SCHL <fct> "1 or more years of college credit, no degree", "Ba…  
$ PINCP <dbl> 41000, 50000, 38000, 30000, 0, 94000, 68000, 120000…  
$ ADJINC <dbl> 1042311, 1042311, 1042311, 1042311, 1042311, 104231…  
$ WAGP <dbl> 14000, 50000, 38000, 0, 0, 94000, 68000, 120000, 20…  
$ PWGTP <dbl> 30, 29, 175, 25, 12, 25, 18, 35, 14, 28, 76, 27, 32…  
$ `O\*NET-SOC Title` <chr> "General and Operations Managers", "General and Ope…  
$ `Green Job Flag` <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

To group by the O\*NET-SOC Sector and demographic factors like education, race, gender, and income, we can create summary statistics for each sector to analyze how the green jobs are distributed across different demographic categories.

# Group the data by sector and demographic variables, then summarize the counts and income  
green\_job\_summary <- merged\_green\_jobs\_data %>%  
 group\_by(`O\*NET-SOC Sector`, SCHL, RAC1P, SEX) %>%  
 summarise(  
 total\_jobs = n(), # Count total jobs  
 mean\_income = mean(PINCP, na.rm = TRUE) # Calculate mean income for each group  
 )

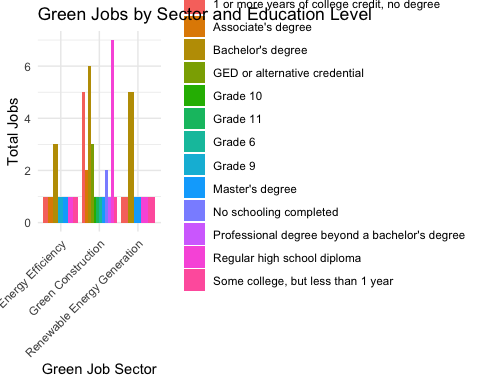
`summarise()` has grouped output by 'O\*NET-SOC Sector', 'SCHL', 'RAC1P'. You  
can override using the `.groups` argument.

# View the summarized data  
print(green\_job\_summary)

# A tibble: 52 × 6  
# Groups: O\*NET-SOC Sector, SCHL, RAC1P [46]  
 `O\*NET-SOC Sector` SCHL RAC1P SEX total\_jobs mean\_income  
 <fct> <fct> <fct> <fct> <int> <dbl>  
 1 Energy Efficiency 1 or more years of col… Whit… Fema… 1 15000   
 2 Energy Efficiency 1 or more years of col… Whit… Male 1 41000   
 3 Energy Efficiency Associate's degree Whit… Male 1 46500   
 4 Energy Efficiency Bachelor's degree Two … Fema… 1 102000   
 5 Energy Efficiency Bachelor's degree Whit… Fema… 2 40000   
 6 Energy Efficiency Bachelor's degree Whit… Male 3 126417.  
 7 Energy Efficiency Grade 9 Whit… Fema… 1 0   
 8 Energy Efficiency Master's degree Asia… Male 1 80000   
 9 Energy Efficiency Master's degree Whit… Fema… 1 113250   
10 Energy Efficiency Regular high school di… Whit… Male 1 94000   
# ℹ 42 more rows

Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

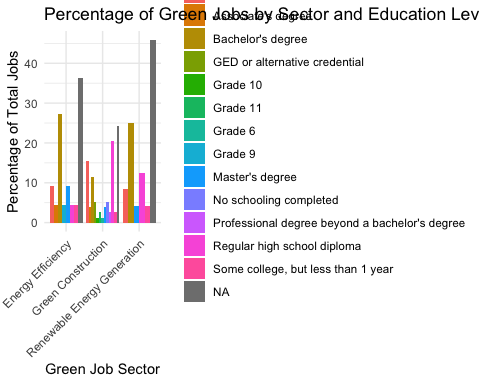
# Remove rows with NA in 'SCHL' or 'O\*NET-SOC Sector'  
cleaned\_green\_job\_summary <- green\_job\_summary %>%  
 filter(!is.na(SCHL), !is.na(`O\*NET-SOC Sector`))  
  
# Bar plot for green jobs by sector and education level  
ggplot(cleaned\_green\_job\_summary, aes(x = `O\*NET-SOC Sector`, y = total\_jobs, fill = SCHL)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 labs(title = "Green Jobs by Sector and Education Level",  
 x = "Green Job Sector", y = "Total Jobs") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



# Plot percentage of jobs by sector and education level  
green\_job\_summary\_percentage <- merged\_green\_jobs\_data %>%  
 group\_by(`O\*NET-SOC Sector`, SCHL) %>%  
 summarise(total\_jobs = n()) %>%  
 mutate(percentage\_jobs = total\_jobs / sum(total\_jobs) \* 100)

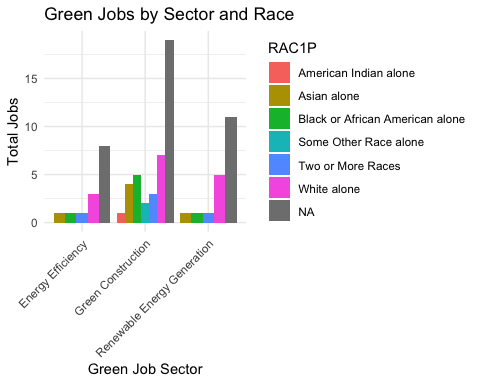
`summarise()` has grouped output by 'O\*NET-SOC Sector'. You can override using  
the `.groups` argument.

ggplot(green\_job\_summary\_percentage, aes(x = `O\*NET-SOC Sector`, y = percentage\_jobs, fill = SCHL)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 labs(title = "Percentage of Green Jobs by Sector and Education Level",  
 x = "Green Job Sector", y = "Percentage of Total Jobs") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

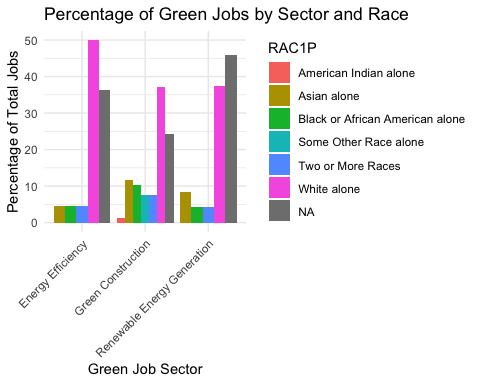
# Bar plot for green jobs by sector and race  
ggplot(green\_job\_summary, aes(x = `O\*NET-SOC Sector`, y = total\_jobs, fill = RAC1P)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 labs(title = "Green Jobs by Sector and Race",  
 x = "Green Job Sector", y = "Total Jobs") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



# Bar plot for green jobs by sector and race (percentage)  
green\_job\_race\_percentage <- merged\_green\_jobs\_data %>%  
 group\_by(`O\*NET-SOC Sector`, RAC1P) %>%  
 summarise(total\_jobs = n()) %>%  
 mutate(percentage\_jobs = total\_jobs / sum(total\_jobs) \* 100)

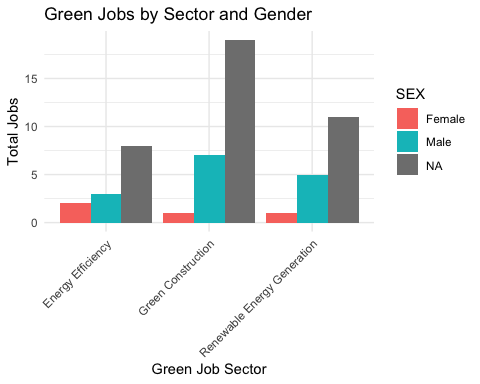
`summarise()` has grouped output by 'O\*NET-SOC Sector'. You can override using  
the `.groups` argument.

ggplot(green\_job\_race\_percentage, aes(x = `O\*NET-SOC Sector`, y = percentage\_jobs, fill = RAC1P)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 labs(title = "Percentage of Green Jobs by Sector and Race",  
 x = "Green Job Sector", y = "Percentage of Total Jobs") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

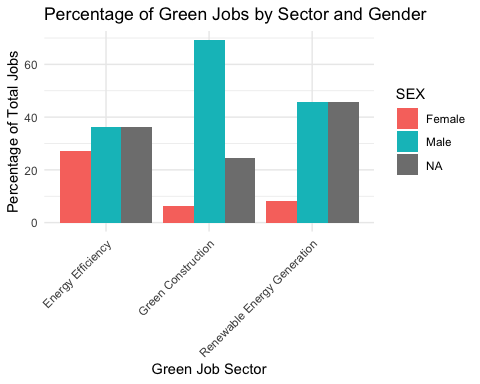
# Bar plot for green jobs by sector and gender  
ggplot(green\_job\_summary, aes(x = `O\*NET-SOC Sector`, y = total\_jobs, fill = SEX)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 labs(title = "Green Jobs by Sector and Gender",  
 x = "Green Job Sector", y = "Total Jobs") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



# Bar plot for green jobs by sector and gender (percentage)  
green\_job\_gender\_percentage <- merged\_green\_jobs\_data %>%  
 group\_by(`O\*NET-SOC Sector`, SEX) %>%  
 summarise(total\_jobs = n()) %>%  
 mutate(percentage\_jobs = total\_jobs / sum(total\_jobs) \* 100)

`summarise()` has grouped output by 'O\*NET-SOC Sector'. You can override using  
the `.groups` argument.

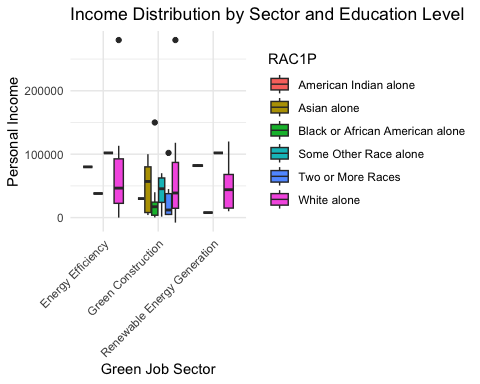
ggplot(green\_job\_gender\_percentage, aes(x = `O\*NET-SOC Sector`, y = percentage\_jobs, fill = SEX)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 labs(title = "Percentage of Green Jobs by Sector and Gender",  
 x = "Green Job Sector", y = "Percentage of Total Jobs") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)

# Box plot for income distribution by sector and education level  
ggplot(merged\_green\_jobs\_data, aes(x = `O\*NET-SOC Sector`, y = PINCP, fill = RAC1P)) +  
 geom\_boxplot() +  
 labs(title = "Income Distribution by Sector and Education Level",  
 x = "Green Job Sector", y = "Personal Income") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))

Warning: Removed 38 rows containing non-finite outside the scale range  
(`stat\_boxplot()`).



Source: [Article Notebook](https://beeckcenter.github.io/climate-equity-workforce/index-preview.html)