Week 3 Plots, Tables, and Stories Work

Naren Prakash

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(plotly)
```

```
Attaching package: 'plotly'

The following object is masked from 'package:ggplot2':

    last_plot

The following object is masked from 'package:stats':

    filter

The following object is masked from 'package:graphics':

    layout
```

```
library(sf)
```

```
Linking to GEOS 3.12.2, GDAL 3.9.3, PROJ 9.4.1; sf_use_s2() is TRUE
```

```
wa_counties <- read_sf("C:/Users/naren/Dropbox/STATS 140XP/Week 3/Plots Tables and Stories/y
ev_data <- read_csv("C:/Users/naren/Dropbox/STATS 140XP/Week 3/Plots Tables and Stories/your
```

```
Rows: 223995 Columns: 17
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (12): VIN (1-10), County, City, State, Postal Code, Make, Model, Electri...
dbl  (5): Model Year, Electric Range, Base MSRP, Legislative District, DOL V...

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.
```

```
demographics <- read_csv("C:/Users/naren/Dropbox/STATS 140XP/Week 3/Plots Tables and Stories/
```

```
Rows: 39 Columns: 31
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (1): County
dbl (30): Pop_18_24, Pop_18_24_Less_than_HS_grad, Pop_18_24_HS_grad_includes...

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.
```

Visual idea: Find outlier counties in Seattle by earnings (after combining age groups), take median earnings, then compare those earnings to the distribution of the medians of all the counties.

Focusing on population over 25

```
print(colnames(ev_data))
```

```
 [1] "VIN (1-10)"
 [2] "County"
 [3] "City"
 [4] "State"
```

```
 [5] "Postal Code"
 [6] "Model Year"
 [7] "Make"
 [8] "Model"
 [9] "Electric Vehicle Type"
[10] "Clean Alternative Fuel Vehicle (CAFV) Eligibility"
[11] "Electric Range"
[12] "Base MSRP"
[13] "Legislative District"
[14] "DOL Vehicle ID"
[15] "Vehicle Location"
[16] "Electric Utility"
[17] "2020 Census Tract"
```

```
demo <- demographics %>%
    select(c(County, Pop_25_over, MEDIAN_EARNINGS_2022_25_over))
ev_sub <- ev_data %>%
    select(c(County, `DOL Vehicle ID`))

combined <- full_join(demo, ev_sub)
```

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

```
combined <- combined %>%
    right_join(wa_counties, by = c(County = "JURISDICT_LABEL_NM"))
```

```
print(colnames(combined))
```

```
 [1] "County"                        "Pop_25_over"
 [3] "MEDIAN_EARNINGS_2022_25_over" "DOL Vehicle ID"
 [5] "OBJECTID"                      "JURISDICT_SYST_ID"
 [7] "JURISDICT_TYPE_CD"             "JURISDICT_NM"
 [9] "JURISDICT_DESG_CD"             "JURISDICT_FIPS_DESG_CD"
[11] "JURISDICT_VACATED_FLG"         "EDIT_DATE"
[13] "EDIT_STATUS"                   "EDIT_WHO"
[15] "GLOBALID"                      "geometry"
```

```
combined <- combined %>%
    select(County, Pop_25_over, MEDIAN_EARNINGS_2022_25_over,
        `DOL Vehicle ID`, geometry)
combined$County <- as.factor(combined$County)
```

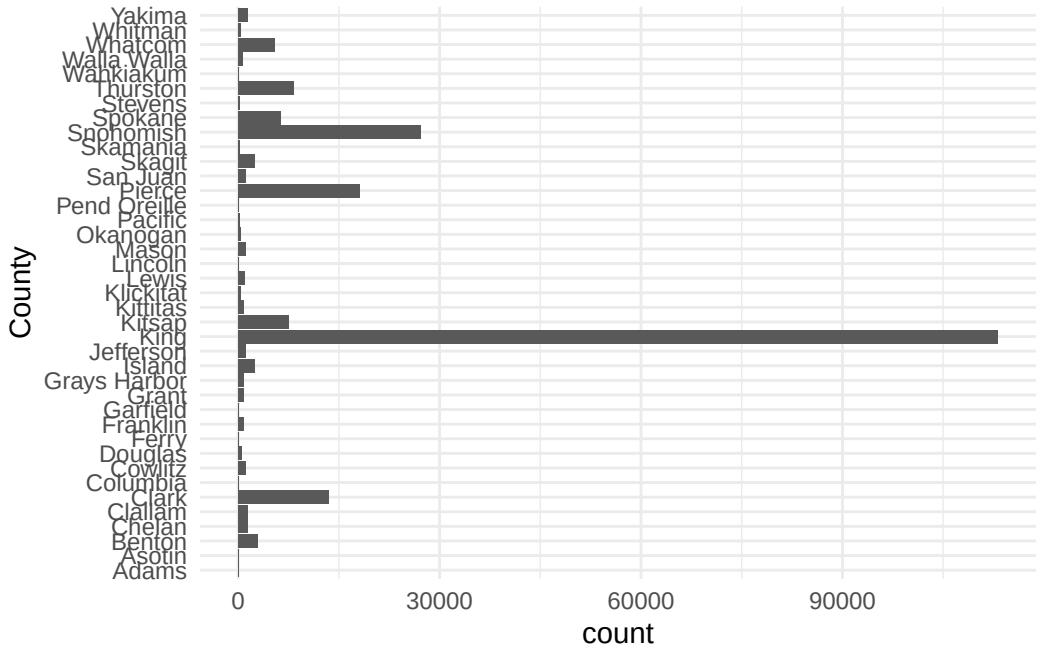
```r
EV_totals <- combined %>%
    group_by(County) %>%
    summarise(EV_count = n()) %>%
    arrange(desc(EV_count))
EV_totals
```

```
# A tibble: 39 x 2
   County    EV_count
   <fct>        <int>
 1 King        113169
 2 Snohomish    27186
 3 Pierce       18026
 4 Clark        13452
 5 Thurston      8252
 6 Kitsap        7476
 7 Spokane       6294
 8 Whatcom       5447
 9 Benton        2892
10 Skagit        2523
# i 29 more rows
```

```r
plot1 <- combined %>%
    ggplot(aes(y = County)) + geom_histogram(stat = "count") +
    theme_minimal()
```

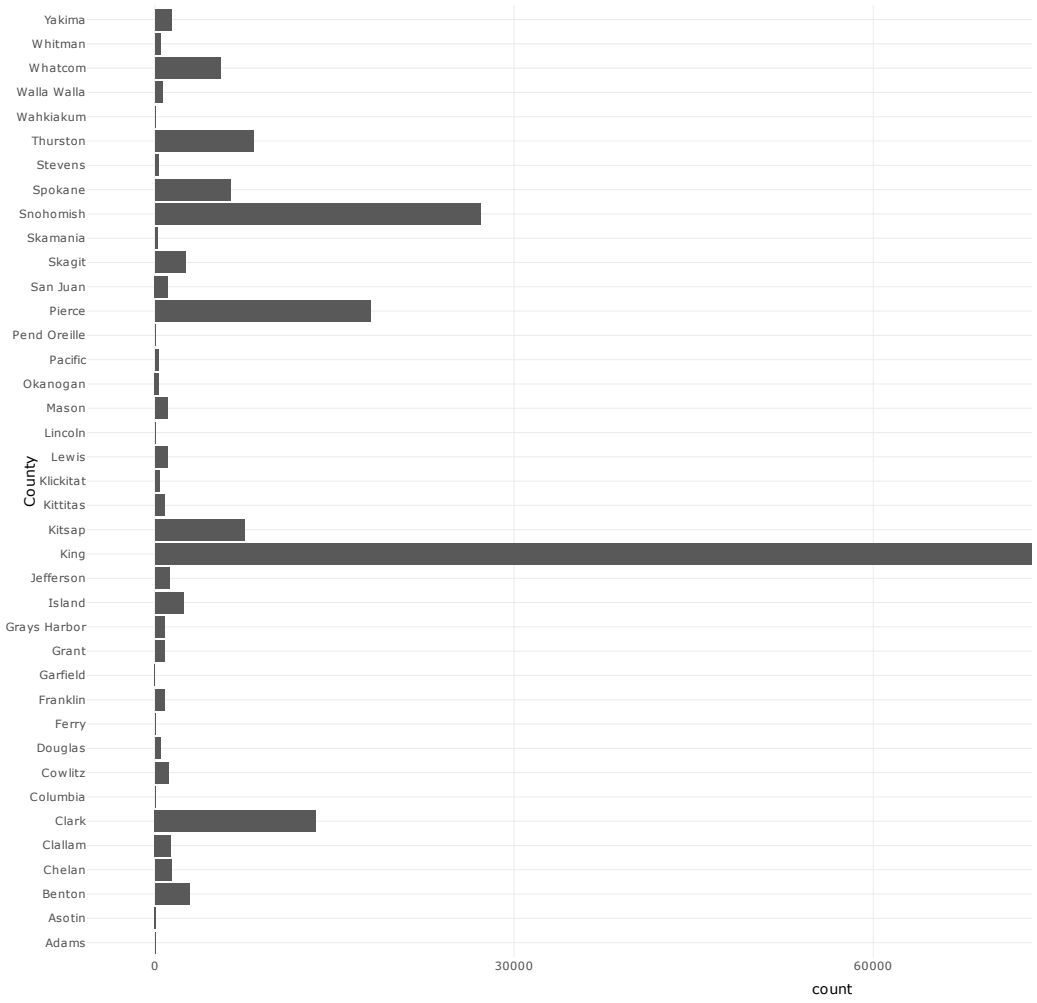
```
Warning in geom_histogram(stat = "count"): Ignoring unknown parameters:
`binwidth`, `bins`, and `pad`
```

```r
plot1
```




```
ggplotly(plot1, show_legend = TRUE)
```

```
file:///C:\Users\naren\AppData\Local\Temp\RtmpW2niFa\file322845c457dd\widget322877aa516b.html
```


Yakima
Whitman
Whatcom
Walla Walla
Wahkiakum
Thurston
Stevens
Spokane
Snohomish
Skamania
Skagit
San Juan
Pierce
Pend Oreille
Pacific
Okanogan
Mason
Lincoln
Lewis
Klickitat
Kittitas
Kitsap
King
Jefferson
Island
Grays Harbor
Grant
Garfield
Franklin
Ferry
Douglas
Cowlitz
Columbia
Clark
Clallam
Chelan
Benton
Asotin
Adams
County
0
30000
60000
count

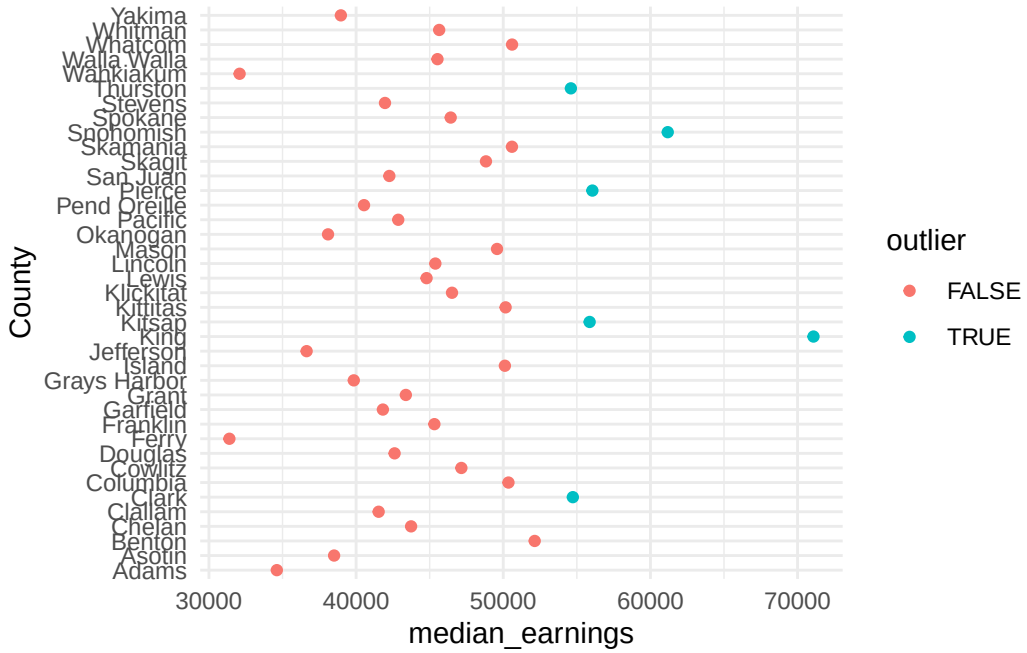
Formally identifying outliers

```
median <- median(EV_totals$EV_count)
iqr <- IQR(EV_totals$EV_count)

EV_totals <- EV_totals %>%
    mutate(scale = (EV_count - median)/iqr)
outlier_df <- EV_totals %>%
    filter(abs(scale) > 3)
outliers <- outlier_df$County
outlier_df
```

```
# A tibble: 6 x 3
  County    EV_count scale
  <fct>        <int> <dbl>
1 King        113169 51.3
2 Snohomish    27186 12.0
3 Pierce       18026  7.86
4 Clark        13452  5.77
5 Thurston      8252  3.39
6 Kitsap        7476  3.04
```

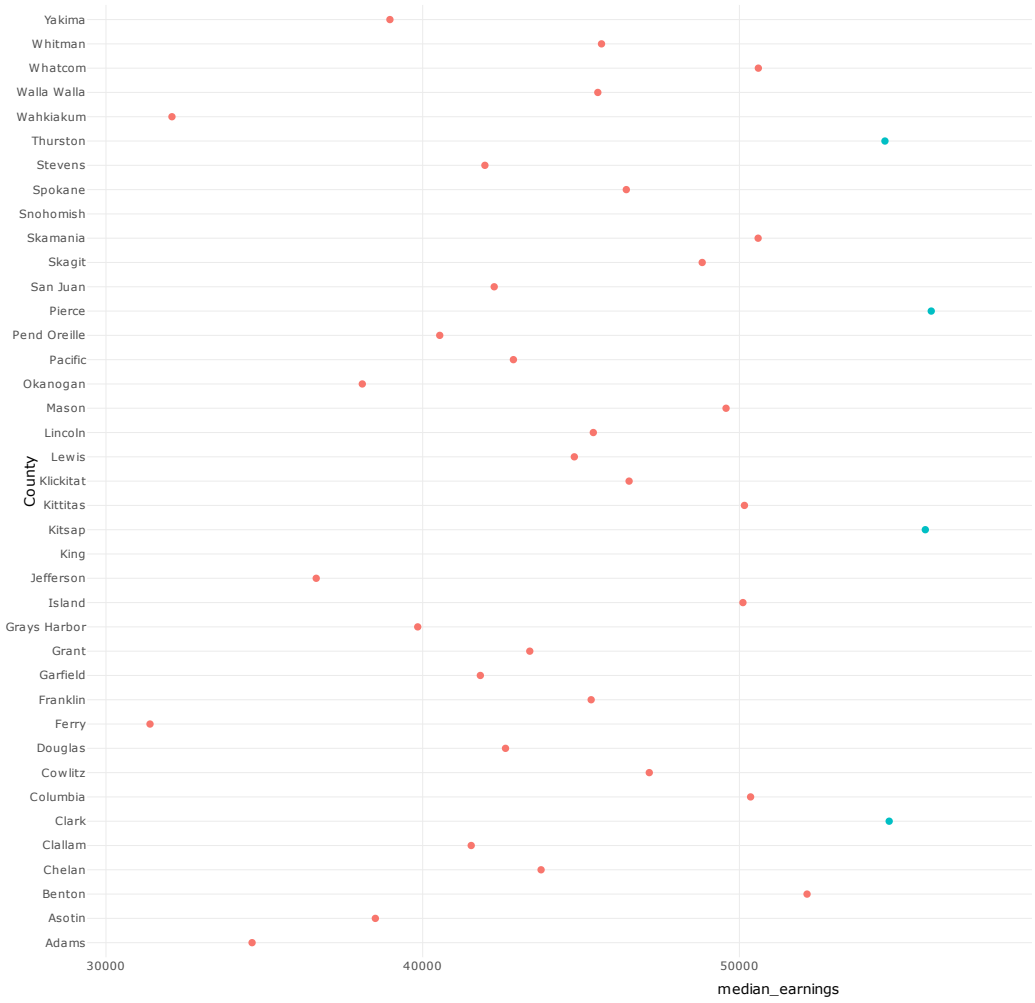
```
county_data <- combined %>%
    group_by(County) %>%
    summarise(median_earnings = mean(MEDIAN_EARNINGS_2022_25_over))
outlier_data <- county_data %>%
    filter(County %in% outliers)

county_data <- county_data %>%
    mutate(outlier = (County %in% outliers))
```

```
plot2 <- county_data %>%
    ggplot(aes(y = County, x = median_earnings, colour = outlier)) +
    geom_point() + theme_minimal()
plot2
```



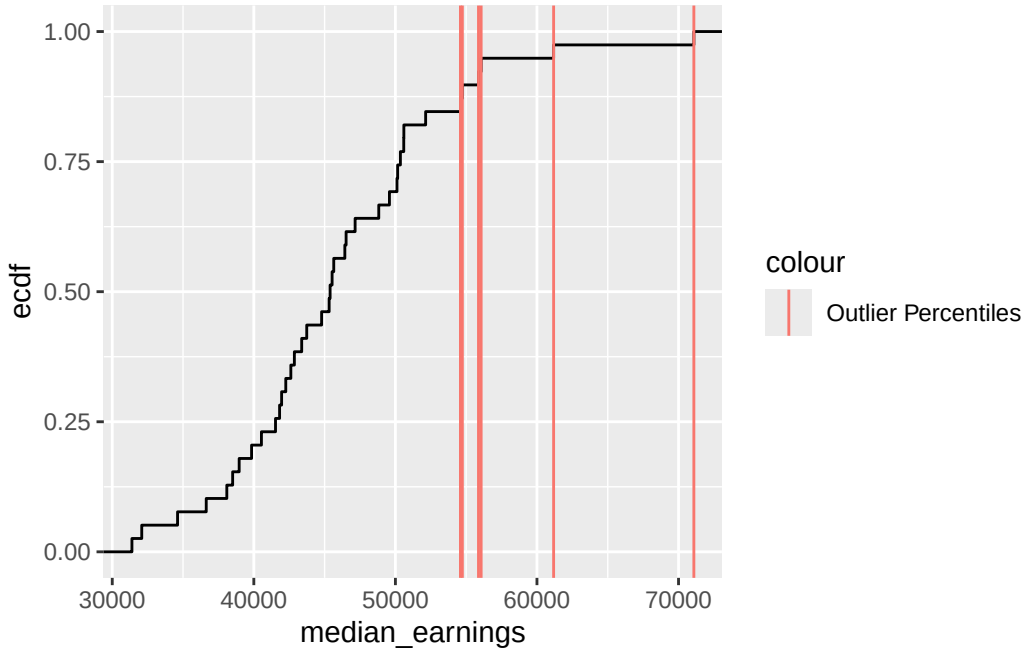

```
ggplotly(plot2, show_legend = TRUE)
```

```
file:///C:\Users\naren\AppData\Local\Temp\RtmpW2niFa\file32283814483c\widget322815dc5926.html
```


Yakima
Whitman
Whatcom
Walla Walla
Wahkiakum
Thurston
Stevens
Spokane
Snohomish
Skamania
Skagit
San Juan
Pierce
Pend Oreille
Pacific
Okanogan
Mason
Lincoln
Lewis
Klickitat
Kittitas
Kitsap
King
Jefferson
Island
Grays Harbor
Grant
Garfield
Franklin
Ferry
Douglas
Cowlitz
Columbia
Clark
Clallam
Chelan
Benton
Asotin
Adams
County
30000
40000
50000
median_earnings

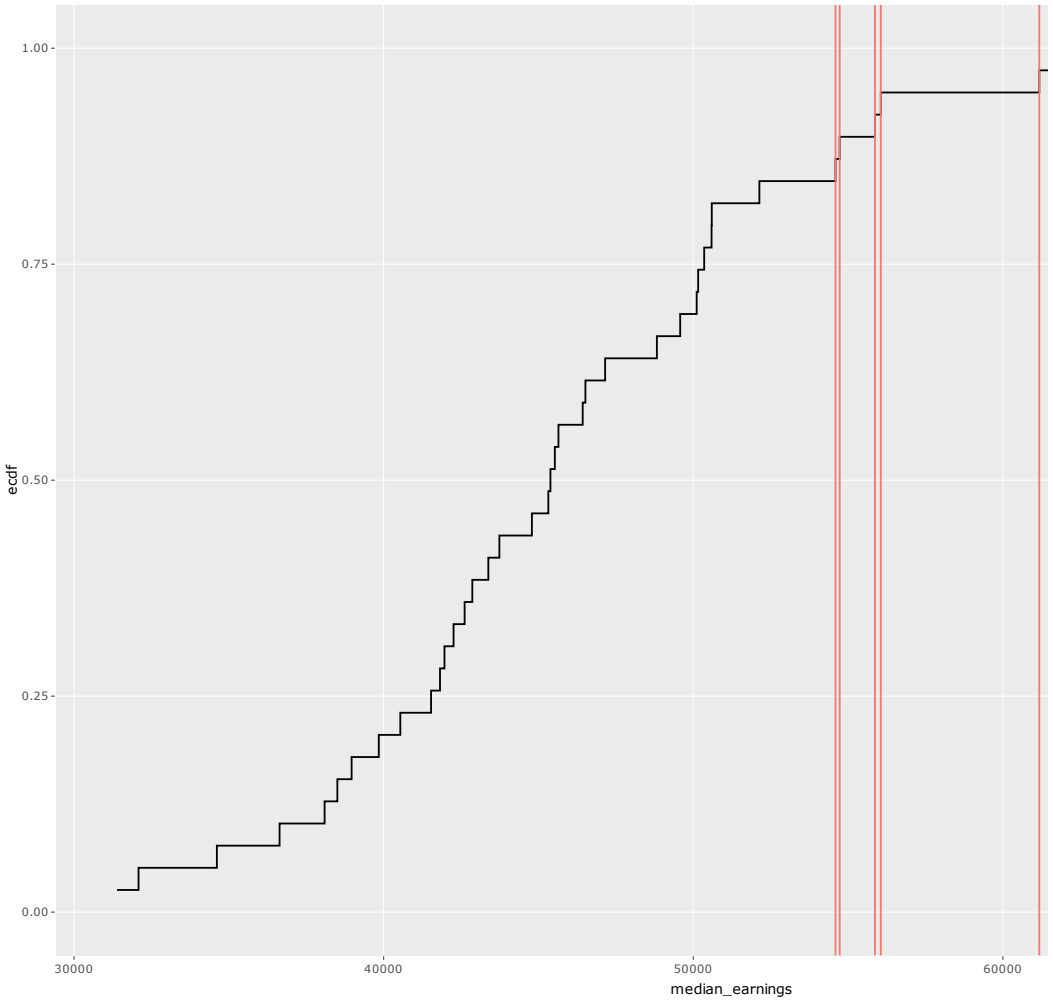
Making a distribution of the medians:

```
p3 <- county_data %>%
    ggplot(aes(x = median_earnings)) + stat_ecdf() + geom_vline(data = outlier_data,
    aes(xintercept = median_earnings, colour = "Outlier Percentiles"))
p3
```




```
ggplotly(p3)
```

```
file:///C:\Users\naren\AppData\Local\Temp\RtmpW2niFa\file322878e33c0d\widget3228510b3790.html
```


1.00
0.75
0.50
0.25
0.00
ecdf
30000
40000
50000
60000
median_earnings

```r
quant <- ecdf(county_data$median_earnings)

county_data <- county_data %>%
    mutate(median_percentile = quant(median_earnings))
```

```r
final <- wa_counties %>%
    left_join(county_data, by = c(JURISDICT_LABEL_NM = "County"))

f <- final %>%
    ggplot(aes(fill = median_percentile)) + geom_sf() + scale_fill_gradient(low = "pink",
    high = "purple") + theme_minimal() + xlab("Longitude") +
    ylab("Latitude") + ggtitle("Median Incomes in Washington State by Percentile")

ggplotly(f, show_legend = TRUE)
```

file:///C:\Users\naren\AppData\Local\Temp\RtmpW2niFa\file3228722546ec\widget322866e437f.html

Median Incomes in Washington State by Percentile