DATASCI 306, Fall 2025, Final Group Project

Group : mention your group names here

Throughout this course, you've dedicated yourself to refining your analytical abilities using R programming language. These skills are highly coveted in today's job market!

Now, for the semester project, you'll apply your learning to craft a compelling `Data Story` that can enrich your portfolio and impress prospective employers. Collaborating with a team (up to 5 members of your choosing), you'll construct a Data Story using the data provided in the `data` folder. This data is downloaded from: https://data.cdc.gov/browse?sortBy=relevance&pageSize=20&q=Adult+Tobacco+Consumption+In+The+U.S.%2C+2000-Present&page=1

Deliverable

1. Requirement-1 (4 pt)

You should show at least 4 steps you adopt to clean and/or transform the dataset. Some of the steps you might take are; merging all the data into one dataframe, converting datatypes, creating additional columns, cleaning column names etc.

```
age <- read_csv("data/Age-Related_Disparities_in_Cigarette_Smoking_Among_Adults_20251101.csv")
```

```
## Rows: 7956 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (8): State, Tobacco Use, Demographic, Comparing (Focus group), Cigarette...
## dbl (1): Year
##
## 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.
```

```
employment <- read_csv("data/Employment-Related_Disparities_in_Cigarette_Smoking_Among_Adults_20251101.
```

```
## Rows: 13260 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (8): State, Tobacco Use, Demographic, Comparing (Focus group), Cigarette...
## dbl (1): Year
##
## 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.
```

```
income <- read_csv("data/Income-Related_Disparities_in_Cigarette_Smoking_Among_Adults_20251101.csv")
```

```
## Rows: 3978 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (8): State, Tobacco Use, Demographic, Comparing (Focus group), Cigarette...
## dbl (1): Year
##
```

```
## 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.
```

```
m_health <- read_csv("data/Mental_Health-Related_Disparities_in_Cigarette_Smoking_Among_Adults_20251101
```

```
## Rows: 3978 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (8): State, Tobacco Use, Demographic, Comparing (Focus group), Cigarette...
## dbl (1): Year
##
## 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.
```

```
race <- read_csv("data/Race_and_Ethnic_Disparities_in_Cigarette_Smoking_Among_Adults_20251101.csv")
```

```
## Rows: 13260 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (8): State, Tobacco Use, Demographic, Comparing (Focus group), Cigarette...
## dbl (1): Year
##
## 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.
```

```
# Merging all tables into one. This only requires bind_rows since all variable names are the same acros

tobacco_chr <- bind_rows(age, employment, income, m_health, race)
```

```
# Converting quantitative variables to numeric datatypes

tobacco <- tobacco_chr |>
  mutate(across(c(`Cigarette Use Prevalence % (Focus group)`,
                  `Cigarette Use Prevalence % (Reference group)`,
                  `Disparity Value`),
                as.numeric))
```

```
## Warning: There were 3 warnings in `mutate()`.
## The first warning was:
## i In argument: `across(...)`.
## Caused by warning:
## ! NAs introduced by coercion
## i Run `dplyr::last_dplyr_warnings()` to see the 2 remaining warnings.
```

```
n_distinct(tobacco$`Tobacco Use`)
```

```
## [1] 1
```

```
# The "tobacco use" column is useless since it only says "Cigarette Use Among Adults". Therefore let's
tobacco <- tobacco |>
  select(-`Tobacco Use`)
```

```
tobacco <- tobacco |>
  group_by(`Year`,
           `State`,
           `Demographic`,
           `Comparing (Focus group)`,
           `Cigarette Use Prevalence % (Focus group)`)|>
```

```
  mutate(`Mean Disparity Value (Focus Group)` = mean(`Disparity Value`))

# Creating a new column which assigns each state to a designated region based on the Census Regions and

tobacco_region <- tobacco |>
  mutate(
    region = case_when(
      State %in% c(
        "Maine","New Hampshire","Vermont","Massachusetts","Rhode Island","Connecticut",
        "New York","New Jersey","Pennsylvania"
      ) ~ "Northeast",

      State %in% c(
        "Ohio","Indiana","Illinois","Michigan","Wisconsin",
        "Minnesota","Iowa","Missouri","North Dakota","South Dakota","Nebraska","Kansas"
      ) ~ "Midwest",

      State %in% c(
        "Delaware","Maryland","District of Columbia","Virginia","West Virginia",
        "North Carolina","South Carolina","Georgia","Florida",
        "Kentucky","Tennessee","Mississippi","Alabama",
        "Oklahoma","Texas","Arkansas","Louisiana"
      ) ~ "South",

      State %in% c(
        "Idaho","Montana","Wyoming","Nevada","Utah","Colorado","Arizona","New Mexico",
        "Alaska","Washington","Oregon","California","Hawaii"
      ) ~ "West",))

tobacco_region
```

```
## # A tibble: 42,432 x 10
## # Groups:   Year, State, Demographic, Comparing (Focus group), Cigarette Use
## #   Prevalence % (Focus group) [13,260]
##     Year State   Demographic `Comparing (Focus group)` Cigarette Use Prevalenc~1
##    <dbl> <chr>   <chr>       <chr>                                          <dbl>
##  1  2011 Alabama Age         Age 18-24                                       30.3
##  2  2011 Alabama Age         Age 18-24                                       30.3
##  3  2011 Alabama Age         Age 18-24                                       30.3
##  4  2011 Alabama Age         Age 25-44                                       28.1
##  5  2011 Alabama Age         Age 25-44                                       28.1
##  6  2011 Alabama Age         Age 25-44                                       28.1
##  7  2011 Alabama Age         Age 45-64                                       26
##  8  2011 Alabama Age         Age 45-64                                       26
##  9  2011 Alabama Age         Age 45-64                                       26
## 10  2011 Alabama Age         Age 65 or older                                 10.2
## # i 42,422 more rows
## # i abbreviated name: 1: `Cigarette Use Prevalence % (Focus group)`
## # i 5 more variables: `To (Reference group)` <chr>,
## #   `Cigarette Use Prevalence % (Reference group)` <dbl>,
## #   `Disparity Value` <dbl>, `Mean Disparity Value (Focus Group)` <dbl>,
## #   region <chr>
```

```r
#To check for NA values.
tobacco_region |>
  filter(is.na(region))
```

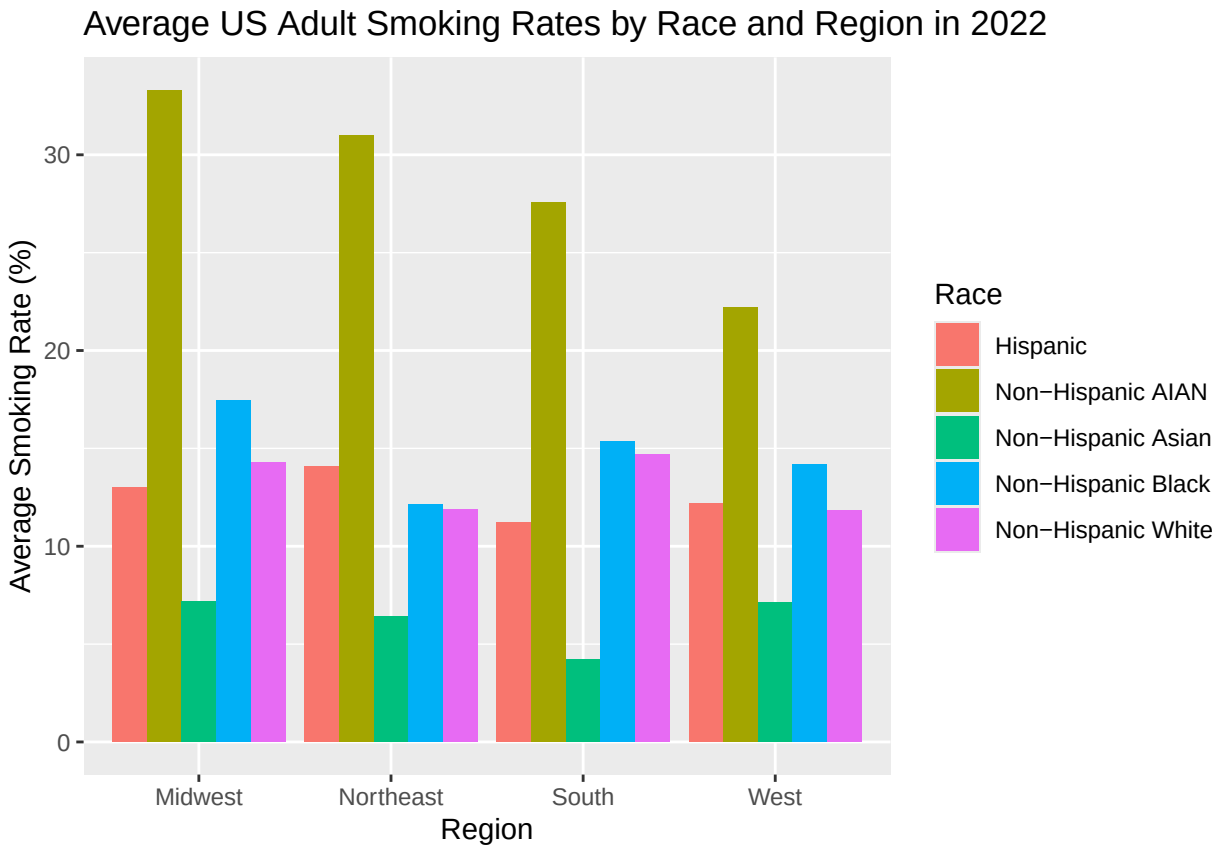
```
## # A tibble: 0 x 10
## # Groups:   Year, State, Demographic, Comparing (Focus group), Cigarette Use
## #   Prevalence % (Focus group) [0]
## # i 10 variables: Year <dbl>, State <chr>, Demographic <chr>,
## #   Comparing (Focus group) <chr>,
## #   Cigarette Use Prevalence % (Focus group) <dbl>, To (Reference group) <chr>,
## #   Cigarette Use Prevalence % (Reference group) <dbl>, Disparity Value <dbl>,
## #   Mean Disparity Value (Focus Group) <dbl>, region <chr>
```

2. Requirement - 2 (20 pt)

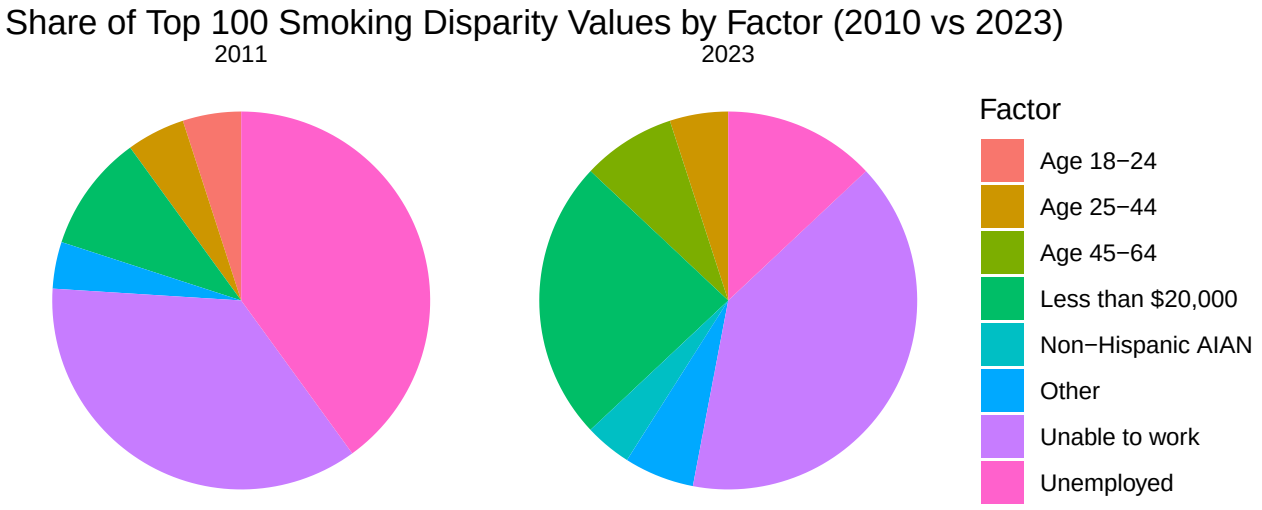
You need to plot 10 different diagrams to show correlations, frequencies, and/or relationships between various variables with plots of 5 different types (bar, line, heatmap, facet, etc.). Every plot should have a title and the x/y axis should have legible labels without any label overlaps for full credit. Provide a summary of your interpretations from the plots after each one. Each chart should be meaningful by itself.

```r
tobacco_region |>
  filter(Year == 2022, Demographic == "Race and Ethnicity") |>

  group_by(Race = `Comparing (Focus group)`, region) |>
  summarize(race_mean = mean(`Cigarette Use Prevalence % (Focus group)`, na.rm = TRUE)) |>

  ggplot(aes(x = region, y = race_mean, fill = Race)) + geom_col(position = "dodge") +
  labs(title = "Average US Adult Smoking Rates by Race and Region in 2022",
       x = "Region", y = "Average Smoking Rate (%)", fill = "Race")
```

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

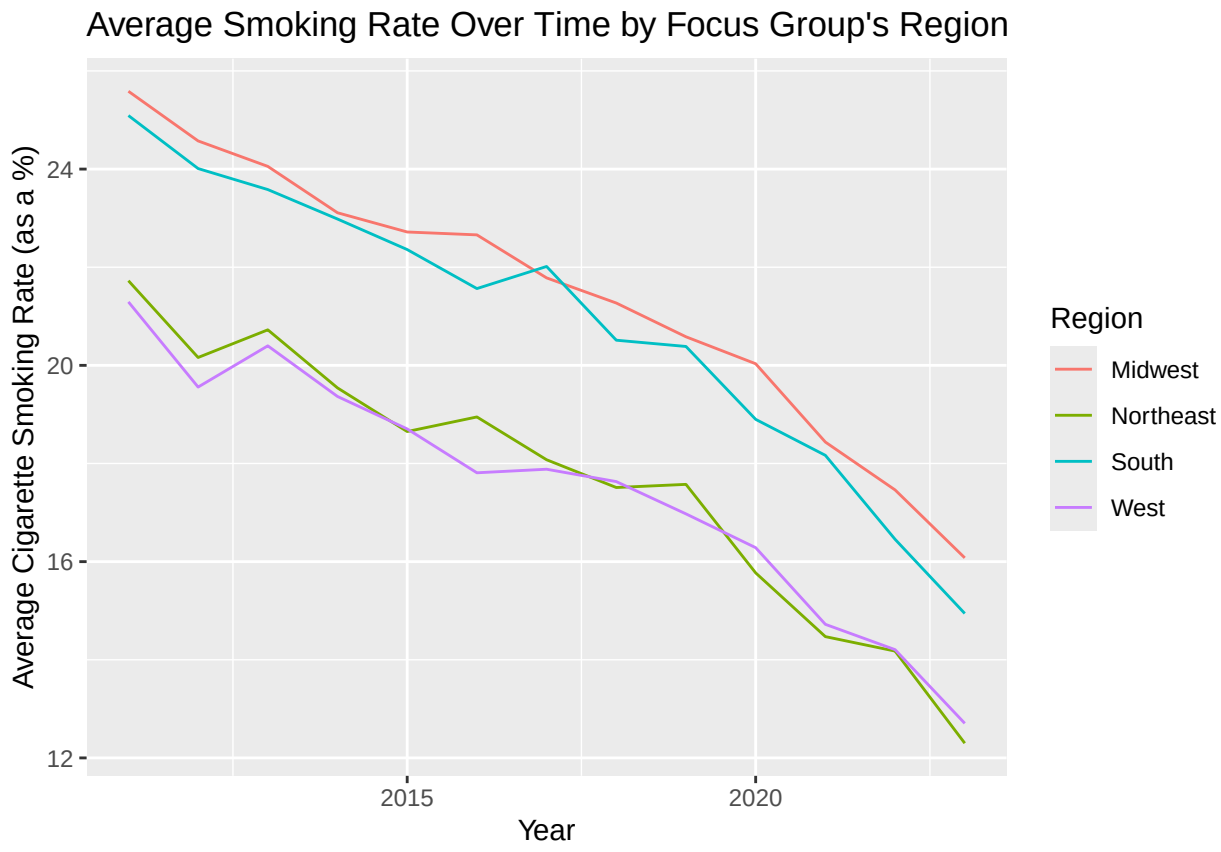



```
tobacco_region |>
  ungroup() |>
  filter(Year %in% c(2011, 2023)) |>
  select(Year, `Comparing (Focus group)`, `Disparity Value`) |>
  drop_na(`Disparity Value`) |>
  group_by(Year) |>
  slice_max(`Disparity Value`, n = 100, with_ties = F) |>
  add_count(`Comparing (Focus group)`, name = "n_fg") |>
  mutate(FocusGroup2 = ifelse(n_fg <= 3,
                              "Other",
                              `Comparing (Focus group)`)) |>
  ggplot(aes(x = "", fill = FocusGroup2)) +
  geom_bar(width = 1) +
  coord_polar(theta = "y") +
  facet_wrap(~ Year) +
  labs(title = "Share of Top 100 Smoking Disparity Values by Factor (2010 vs 2023)",
      x = NULL, y = NULL,
      fill = "Factor") +
  theme_void()
```



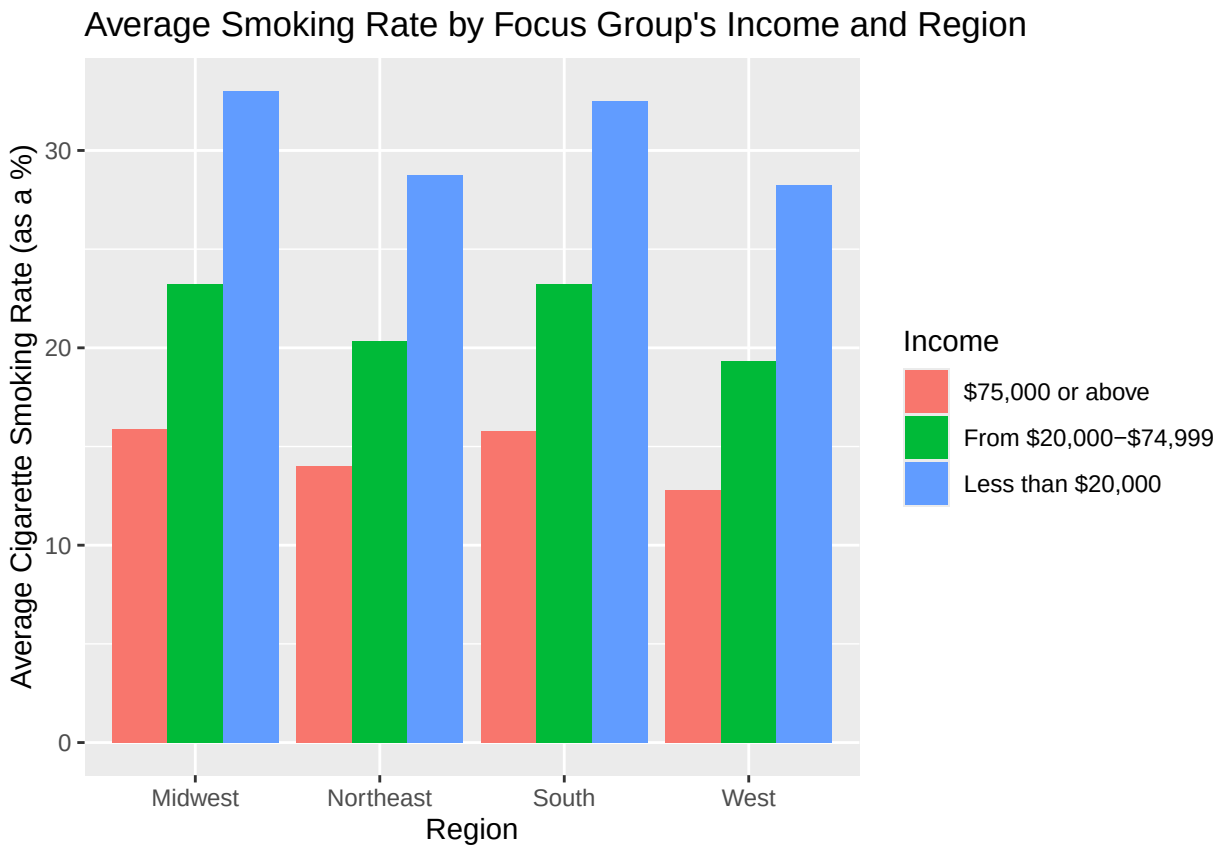

```
tobacco_region |> group_by(`Year`, region) |>
  summarize(avg = mean(`Cigarette Use Prevalence % (Focus group)`, na.rm = T)) |>

  ggplot() + geom_line(aes(x = `Year`,
                           y = avg, color = region)) +
    labs(title = "Average Smoking Rate Over Time by Focus Group's Region",
         y = "Average Cigarette Smoking Rate (as a %)",
         color = "Region")
```

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



```
tobacco_region |> filter(`Demographic` == "Income") |> group_by(`Comparing (Focus group)`, `region`) |>
  summarize(avg = mean(`Cigarette Use Prevalence % (Focus group)`, na.rm = T)) |>

  ggplot(aes(fill = `Comparing (Focus group)`, y = avg, x = region)) + geom_col(position = "dodge") +
    labs(title = "Average Smoking Rate by Focus Group's Income and Region",
         x = "Region",
         y = "Average Cigarette Smoking Rate (as a %)",
         fill = "Income")
```

```
## `summarise()` has grouped output by 'Comparing (Focus group)'. You can override
## using the `.groups` argument.
```




```
# --- Step 1: Convert to numeric ---
m_health_numeric <- m_health %>%
  mutate(across(c(`Cigarette Use Prevalence % (Focus group)`,
                  `Cigarette Use Prevalence % (Reference group)`,
                  `Disparity Value`),
                as.numeric))
```

```
## Warning: There were 3 warnings in `mutate()`.
## The first warning was:
## i In argument: `across(...)`.
## Caused by warning:
## ! NAs introduced by coercion
## i Run `dplyr::last_dplyr_warnings()` to see the 2 remaining warnings.
```

```
income_numeric <- income %>%
  mutate(across(c(`Cigarette Use Prevalence % (Focus group)`,
```

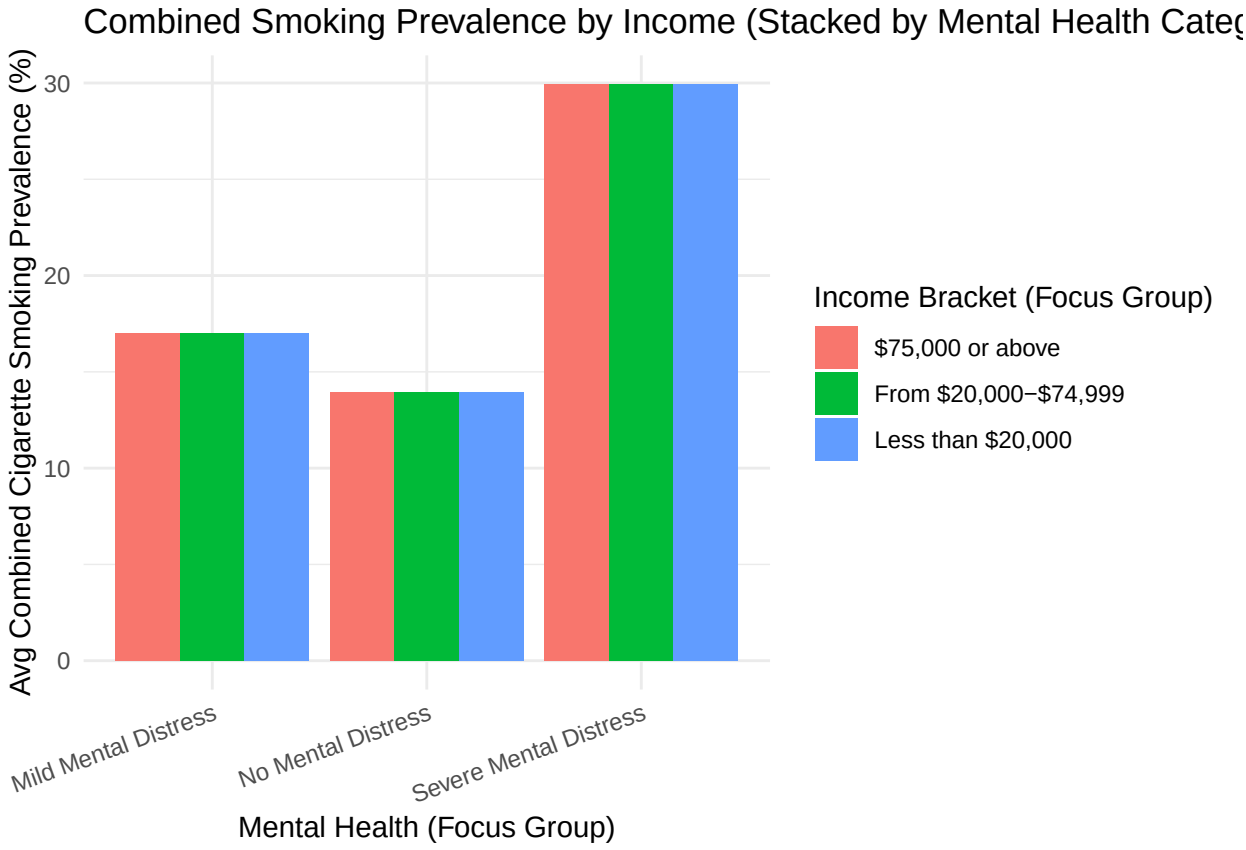
```
                  `Cigarette Use Prevalence % (Reference group)`,
                  `Disparity Value`),
                as.numeric))
```

```
## Warning: There were 3 warnings in `mutate()`.
## The first warning was:
## i In argument: `across(...)`.
## Caused by warning:
## ! NAs introduced by coercion
## i Run `dplyr::last_dplyr_warnings()` to see the 2 remaining warnings.
```

```
# --- Step 2: Clean columns for merging ---
m_health_clean <- m_health_numeric %>%
  select(State, Year,
         MentalHealth = `Comparing (Focus group)`,
         MH_Smoking = `Cigarette Use Prevalence % (Focus group)`)

income_clean <- income_numeric %>%
  select(State, Year,
         Income = `Comparing (Focus group)`,
         Income_Smoking = `Cigarette Use Prevalence % (Focus group)`)

# --- Step 3: Merge income + mental health by state and year ---
merged <- income_clean %>%
  inner_join(m_health_clean, by = c("State", "Year"))
```

```
## Warning in inner_join(., m_health_clean, by = c("State", "Year")): Detected an unexpected many-to-ma
## i Row 1 of `x` matches multiple rows in `y`.
## i Row 1 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.
```

```
# --- Step 4: Compute the COMBINED average smoking prevalence ---
plot_df <- merged %>%
  group_by(Income, MentalHealth) %>%
  summarize(Income_Smoking_avg = mean(Income_Smoking, na.rm = TRUE),
            MH_Smoking_avg = mean(MH_Smoking, na.rm = T),
            .groups = "drop")

# --- Step 5: Stacked bar chart ---
ggplot(plot_df,
       aes(fill = Income,
           y = MH_Smoking_avg,
           x = MentalHealth)) +
  geom_col(position = "dodge") +
  labs(
    title = "Combined Smoking Prevalence by Income (Stacked by Mental Health Category)",
    fill = "Income Bracket (Focus Group)",
    y = "Avg Combined Cigarette Smoking Prevalence (%)",
    x = "Mental Health (Focus Group)"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 20, hjust = 1))
```



3. Requirement - 3 (5 pt)

By this phase, you have a pretty good understanding of your data. Now, you will develop a predictive Machine Learning model to forecast a specified target outcome. For predictor selection, you must apply the feature selection techniques covered in Lectures 19 and 20. Provide a detailed justification for the final set of predictors chosen to receive full credit.

Build an interactive shiny app that allows the user to select or input values and then make predictions using your model. You are required to build the shiny app just for the prediction section of your project. Although we will accept if you create a shiny app for your entire project (not required)

4. Requirement - 4 (1 pt)

You should have a conclusion, highlighting the main insights you were able to derive from your analysis.

This is an open ended project where every team will come up with their own unique insights. We would like to see what each team comes up with.

Submission

- You will upload the zip file containing your final.Rmd, final.pdf files covering the EDA and app.R file for your shiny app, as a deliverable to Canvas
- You will present your findings by creating a video of a maximum of 15 minutes duration, explaining the code and the workings of your project; all team members should explain their part in the project to receive credit. You will share the video URL on Canvas for credit.

It is not necessary to prepare slides (if you do it doesn't hurt) for the presentation. You may speak by showing the diagrams and/or code from your laptop. Every team member should explain their part in the project

along with the insights they derived by explaining the charts and summaries for full credit to each member. An easy way to get this accomplished is to open a Zoom meeting and everyone take turns in explaining while you record the meeting. Add the URL of this recording to Canvas.

Your project will be evaluated for its meaningful/insightful EDA and predictions.

Hints

Some answers you may try to find in this dataset could be:

- Are smoking disparities related to income getting worse or better over time?
- Which racial or ethnic group faces the greatest inequality in smoking rates compared to the majority population?
- Do people with frequent mental distress have a much higher smoking disparity than people with disabilities?
- Which 5 focus groups have the highest average disparity over the entire period?

These questions could just get you started but as an analyst you should hone the skills of asking good questions and this project will get you that practice.

Machine Learning Model Development:

You could build a regression model to predict the DisparityValue. Features may include, Year, disparity category, specific focus group etc. Extract and plot the feature importance to select the features.

Summary

Your summary might answer questions like;

- What were the most significant trends and predictors of smoking disparity?
- Which populations should be prioritized for smoking cessation programs? etc.
  You may also include ideas for future analysis and any limitations you came across with the current dataset