Milestone #3

Rachael Baartmans, Lara Petalio, Christine Truong

Sys.Date()

```
library(readr)
library(tidyverse)
```

```
## Warning in system("timedatectl", intern = TRUE): running command 'timedatectl'
## had status 1
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v dplyr   1.0.8
## v tibble  3.1.6     v stringr 1.4.0
## v tidyr   1.2.0     v forcats 0.5.1
## v purrr   0.3.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
race_data <-read_csv("ca_csc_outcome_race_data.csv",
          col_select = c(NERVOUS, WORRYING, PROBINTR,
                         PROBDOWN, ASTHMA, HEARTDIS,
                         DIABETES, OTHMENILL, race01, race02, race03,
                         race04, race05, race06, race07, race08,
                         race09, race10, race11, race12, race13,
                         race14, race15),
          na = c("", "NA", "NA/Not Applicable", "N/A", "n/a",
                 "(DO NOT READ) NA/Not Applicable",
                 "(DO NOT READ) Refused",
                 "(DO NOT READ) Don't know"))
```

```
## Rows: 1000 Columns: 23
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (23): NERVOUS, WORRYING, PROBINTR, PROBDOWN, ASTHMA, HEARTDIS, DIABETES,...
##
## i Use 'spec()' to retrieve the full column specification for this data.
## i Specify the column types or set 'show_col_types = FALSE' to quiet this message.
```

```r
smoker_data <- read_csv("ca_csc_smoker_data.csv",
          col_select = c(smokstat, WHEREBUY, BUYCALIF,
                         HOWMANY, SMOK6NUM, SMOK6UNI),
          na = c("", "NA", "NA/Not Applicable", "N/A", "n/a",
                 "(DO NOT READ) NA/Not Applicable",
                 "(DO NOT READ) Refused",
                 "(DO NOT READ) Don't know"))
```

```
## Rows: 1000 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): smokstat, HOWMANY, SMOK6UNI, BUYCALIF, WHEREBUY
## dbl (1): SMOK6NUM
##
## 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.
```

Subset rows and columns as needed

Create New Variables needed for analysis

Minimum of 2 created from existing columns
Examples: calculating the rate or combining character strings

Clean variables for analysis

Minimum of 2
**Examples: Recode invalid values/handle missing fields/recode categories **

Data dictionary based on clean dataset

must include: variable name, data type, and description

Tables with descriptive statistics for 4 data elements