data_cleaning

Betsy Norwood

1/30/2023

Loading in the data

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2
## ──
```

```
## ✔ ggplot2 3.4.0     ✔ purrr   0.3.4
## ✔ tibble  3.1.6     ✔ stringr 1.4.0
## ✔ tidyr   1.1.4     ✔ forcats 0.5.1
## ✔ readr   2.1.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
usa_lung_cancer <- read_csv('Desktop/710Project/usa_lung_cancer.csv', skip=1)
```

```
## Rows: 3234 Columns: 10
## ── Column specification ────────────────────────────────────────────────────
## Delimiter: ","
## chr (9): GeoID_Description, GeoID_Name, SitsinState, GeoID, GeoID_Formatted,...
## dbl (1): GeoVintage
##
## 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.
```

```
View(usa_lung_cancer)
```

Renaming columns

```
usa_lung_cancer <- usa_lung_cancer %>% rename_at('SitsinState', ~'State')

usa_lung_cancer <- usa_lung_cancer %>% rename_at('GeoID_Name', ~'County')

usa_lung_cancer <- usa_lung_cancer %>% rename_at('r_l_allr_u_alla', ~'lung_cancer_per
_100000')
```

Removing columns

```
usa_lung_cancer <- usa_lung_cancer %>% select(-one_of('GeoID_Formatted', 'Source', 'L
ocation', 'GeoID_Description', 'GeoVintage'))
View(usa_lung_cancer)
```