Milestone 3 Workflow

Meredith Klashhman and Ruben Vargas

11/8/2021

Connect to Data Files

```
getwd()
```

```
## [1] "/home/rstudio/PHW251_MK_RV"
```

```
countydemo <- read.csv("/home/rstudio/PHW251_MK_RV/ca_county_demographics.csv",
                       header=TRUE)
mortality <- read.csv("/home/rstudio/PHW251_MK_RV/mort_by_county.csv", header =
                        TRUE)
hospital_cost <- read.csv("/home/rstudio/PHW251_MK_RV/oshpd_hospital_cost.csv",
                          header = TRUE)
```

Descriptive Statistics to Check Data Elements

Check Mortality Counts for statistics and NA values

```
summary(as.numeric(mortality$Count))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0     0.0    14.0   266.9    93.0 65404.0   38849
```

```
sum(is.na(mortality$Count))
```

```
## [1] 38849
```

Will need to replace NA values in this dataset.

Check Demograpihcs Data for Renter and Owner Outliers

```
summary(as.numeric(countydemo$renter_occ))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     140    6080   25140   95554   84189 1696455
```

```
summary(as.numeric(countydemo$owner_occ))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     357   13089   39306  121300  120804 1544749
```

Quarter values and min/max indicate good distribution, no need for outlier exclusion.

Check dates in mortality dataset for outliers

```
summary(as.numeric(mortality$Year))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2014    2015    2016    2016    2018    2019
```

No dates are missing, and all fall within a 5 year range. Data is quite clean!

Mortality Data Edits

DATA CLEANING

```
#Edit Mortality to replace NA with 0

mortality[is.na(mortality)] = 0
```

NEW VARIABLE

```
#Create a Variable to Indicate Chronic Condition

#Chronic Health Conditions defined as the following:
    #Alzheimer's disease
    #Chronic lower respiratory diseases
    #Diabetes mellitus
    #Essential hypertension and hypertensive renal disease
    #Chronic liver disease and cirrhosis
    #Parkinson's disease

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
```

```
mortality <- mortality %>%
  mutate(Chronic = case_when(Cause_Desc %in% c("Alzheimer's disease",
                    "Chronic lower respiratory diseases", "Diabetes mellitus",
                    "Essential hypertension and hypertensive renal disease",
                    "Chronic liver disease and cirrhosis",
                    "Parkinson's disease") ~ "Yes",
                    TRUE ~ "No"))
```

SUBSET DATA

```
#subtotal for mortality from chronic in each county per year

library(tidyr)
mortality_chronic <- mortality %>%
```

```
            filter(Chronic == "Yes")%>%
            filter(Strata == "Total Population") %>%
            group_by(County,Year) %>%
            mutate(County_Year_Chronic_Mortality = sum(Count))
```

Demographics Data Edits

NEW VARIABLE

```
#Calculate Rent vs Homeowners Ratio

library(dplyr)

countydemo <- countydemo %>%
              mutate(RatioRenttoOwn = renter_occ/owner_occ) %>%
              mutate (County = name)
```

SUBSET DATA

```
#locate 5 counties that share three common attributes: low population per
#square mile 'pop12_sqmi1', high median age
#'med_age', a high proportion of renters vs. homeowners

countydemo5 <- countydemo %>%  filter (pop12_sqmi < 30) %>%
                filter (med_age > 30) %>%
                filter (RatioRenttoOwn > 0.6)

print(countydemo5$name)
```

```
## [1] "Mendocino" "Mono"      "Colusa"    "Del Norte" "Glenn"
```

Hospital Closure Data Edits

SUBSET DATA

```
#locate the most recent account of OSHPD funding for projects that are in
#closure

hospital_cost_closure <- hospital_cost %>% filter(OSHPD.Project.Status ==
                                                    "In Closure")
```

DATA CLEANING

```
#Optimize data for joining (edit so all county naming is same format)

library(tidyr)

hospital_cost_closure_clean <- hospital_cost_closure %>%
                       separate(County, c('Number','County_Name_1',
                       'County_Name_2', 'County_Name_3' ))
```

```
## Warning: Expected 4 pieces. Missing pieces filled with `NA` in 11856 rows [1, 2,
## 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

```
hospital_cost_closure_clean [is.na(hospital_cost_closure_clean)] <- ""

hospital_cost_closure_clean <- hospital_cost_closure_clean %>% unite(County,
                              'County_Name_1','County_Name_2','County_Name_3',
                              sep = " ", remove = TRUE )
```

Data Dictionary for cleaned and newly created variables

Count

Found in the dataset 'mortality' and 'mortality_chronic'. In both datasets this value indicates the total number of deaths attributed to that year, county, strata, and cause.

```
class(mortality$Count)
```

```
## [1] "numeric"
```

Chronic

Found in the dataset 'mortality' and 'mortality_chronic'. In both datasets this value indicates whether the indicated disease category is considered a chronic condition.

```
class(mortality$Chronic)
```

```
## [1] "character"
```

County_Year_Chronic_Mortality

In the 'mortality_chronic' dataset, this value indicates the number of total chronic deaths attributed to that year and county.

```
class(mortality_chronic$County_Year_Chronic_Mortality)
```

```
## [1] "numeric"
```

County

In the 'hospital_cost_closure_clean' dataset, this value indicates the County, and its data is formatted consistently with the other two datasets, which should allow for easy merging.

```
class(hospital_cost_closure_clean$County)
```

```
## [1] "character"
```