R Getting Started

```
install.packages('tidyverse')
```

```
library('tidyverse')
```

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

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

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

Read in data Often the first step is to connect to the data:

```
con <- DBI::dbConnect(RMySQL::MySQL(),
  host = "database.company.com",
  user = "user_name",
  password = rstudioapi::askForPassword("Database password")
)

A_DATA_TBL <- tbl(con, "table_name")
```

```
A_DATA <- readRDS('A_DATA.RDS')
A_DATA <- readxl::read_excel('A_DATA.xlsx') # don't forget about sheets!
```

```
A_DATA <- readr::read_csv('A_DATA.csv')
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   Gender = col_character(),
##   Race = col_character(),
##   USAF = col_character(),
##   Birth_Country = col_character(),
##   Grade_Level = col_character(),
##   Grade_Range = col_character(),
##   Marital_Status = col_character(),
##   Pregnant = col_character(),
##   Household_Icome = col_logical(),
##   Family_Income = col_logical(),
##   yr_range = col_character(),
##   BMIHEAD = col_logical(),
##   BMAUPREL = col_logical(),
##   BMAUPLEL = col_logical(),
##   BMALLKNE = col_logical(),
##   BMDSTATS = col_logical(),
##   BMDRECUF = col_logical(),
##   BMDSUBF = col_logical(),
##   BMDTHICF = col_logical(),
##   BMDLEGF = col_logical()
##   # ... with 46 more columns
## )
## i Use `spec()` for the full column specifications.
```

Transform and Summarise Data Use dplyr and the tidyverse to get summary data

```
A_DATA %>%
  mutate(over_50 = if_else(Age > 50, 1, 0)) %>%
  group_by(DIABETES, over_50) %>%
  summarise(n = n(),
            mean_age = mean(Age, na.rm = TRUE),
            .groups = 'keep')
```

```
## # A tibble: 6 x 4
## # Groups:   DIABETES, over_50 [6]
##   DIABETES over_50     n mean_age
##      <dbl>   <dbl> <int>    <dbl>
## 1        0       0 68703    19.5
## 2        0       1 20037    66.3
## 3        1       0  1373    38.8
## 4        1       1  5434    67.3
## 5       NA       0  4938     2.77
## 6       NA       1   831    66.8
```

```
A_DATA %>%
  mutate(over_50 = if_else(Age > 50, 1, 0)) %>%
  group_by(DIABETES, over_50) %>%
  summarise(n = n(),
            mean_age = mean(Age, na.rm = TRUE),
            .groups = 'keep') %>%
  pivot_longer(cols=c(n,mean_age),
               names_to = "statistic",
               values_to = "value")
```

```
## # A tibble: 12 x 4
## # Groups:   DIABETES, over_50 [6]
##    DIABETES over_50 statistic    value
##       <dbl>   <dbl> <chr>        <dbl>
##  1        0       0 n          68703
##  2        0       0 mean_age     19.5
##  3        0       1 n          20037
##  4        0       1 mean_age     66.3
##  5        1       0 n           1373
##  6        1       0 mean_age     38.8
##  7        1       1 n           5434
##  8        1       1 mean_age     67.3
##  9       NA       0 n           4938
## 10       NA       0 mean_age      2.77
## 11       NA       1 n            831
## 12       NA       1 mean_age     66.8
```

```
A_DATA %>%
  filter(!is.na(Grade_Level)) %>%
  group_by(Gender, Race) %>%
  tally()
```

```
## # A tibble: 10 x 3
## # Groups:   Gender [2]
##    Gender Race                n
##    <chr>  <chr>           <int>
##  1 Female Black            4084
##  2 Female Mexican American 4143
##  3 Female Other            1356
##  4 Female Other Hispanic   1138
##  5 Female White            3899
##  6 Male   Black            4195
##  7 Male   Mexican American 4055
##  8 Male   Other            1364
##  9 Male   Other Hispanic   1152
## 10 Male   White            4089
```

```
A_DATA %>%
  filter(!is.na(Grade_Level)) %>%
  group_by(Gender, Race) %>%
  tally() %>%
  ungroup() %>%
  pivot_wider(Race, names_from= Gender, values_from = n)
```

```
## # A tibble: 5 x 3
##   Race             Female  Male
##   <chr>             <int> <int>
## 1 Black              4084  4195
## 2 Mexican American   4143  4055
## 3 Other              1356  1364
## 4 Other Hispanic     1138  1152
## 5 White              3899  4089
```

Export Data Write and Save Data

```
A_DATA %>% saveRDS('A_DATA.RDS')
A_DATA %>% openxlsx::write.xlsx('A_DATA.xlsx')
A_DATA %>% readr::write_csv('A_DATA.csv')
```

Stats
t-test
ks-test
chi-square test
anova
Models
glm