Tutorial 5 Worksheet AY 21/22 Sem 1

Read your data into R as absent and identify the number of couriers in the dataset.

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

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

```
## v ggplot2 3.3.5     v dplyr   1.0.7
## v tibble  3.1.4     v stringr 1.4.0
## v tidyr   1.1.3     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()
```

```
absent <- read_delim("data/Absenteeism_at_work.csv")
```

```
## Rows: 740 Columns: 21
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ";"
## dbl (21): ID, Reason for absence, Month of absence, Day of the week, Seasons...
```

```
##
## 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.
```

```
absent_count<-select(absent,ID) %>%
   unique() %>%
   count()
absent_count
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1    36
```

1.There are 36 couriers in the dataset.

2.Max absences Obtain the longest 7 absences (with ties equals TRUE). Keep only the courier ID, reason for absence and the Absenteeism time. Make a mental note of the couriers you observe. Courier IDs-14,36,9,28,9,11,36

```
max_absence_tbl<-select (absent , "ID", "Reason for absence" , "Absenteeism time in hours" )
max_absence_tbl<-slice_max(max_absence_tbl,order_by=max_absence_tbl$`Absenteeism time in hours`,n=7)
max_absence_tbl
```

```
## # A tibble: 9 x 3
##      ID `Reason for absence` `Absenteeism time in hours`
##   <dbl>                <dbl>                        <dbl>
## 1    14                   11                          120
```

```
## 2    36                  13                       120
## 3     9                   6                       120
## 4    28                   9                       112
## 5     9                  12                       112
## 6    11                  19                       104
## 7    36                  13                        80
## 8    14                  18                        80
## 9    13                  13                        80
```

Unknown reason for absence

Extract the rows corresponding to reason for absence equals to 0. Keep only the ID, reason for absence, discipline record and absenteeism time. What do you observe? Courier IDs of first 10 - 36,20,28,11,26,13,36,2,7,18. Courier 36,28,11, were in the list of the longest 7 absences, and also have corresponding readon for absence equals to 0. Maybe there is a correlation.

```
reason_tbl<-select(absent, "ID","Reason for absence","Disciplinary failure","Absenteeism time in hours")
reason_tbl<-filter(reason_tbl, reason_tbl$`Reason for absence`==0)
reason_tbl
```

```
## # A tibble: 43 x 4
##       ID `Reason for absence` `Disciplinary failure` `Absenteeism time in hours`
##    <dbl>                <dbl>                   <dbl>                        <dbl>
##  1    36                    0                       1                            0
##  2    20                    0                       1                            0
##  3    28                    0                       1                            0
##  4    11                    0                       1                            0
##  5    36                    0                       1                            0
##  6    13                    0                       1                            0
##  7    36                    0                       1                            0
##  8     2                    0                       1                            0
##  9     7                    0                       1                            0
## 10    18                    0                       1                            0
## # ... with 33 more rows
```

Remove disciplinary failure Remove the rows corresponding to disciplinary failure equals to 1.

```
absent <- filter(absent, absent$`Disciplinary failure` !=1)
absent
```

```
## # A tibble: 700 x 21
##       ID `Reason for absence` `Month of absence` `Day of the week` Seasons
##    <dbl>                <dbl>               <dbl>             <dbl>   <dbl>
##  1    11                   26                   7                 3       1
##  2     3                   23                   7                 4       1
##  3     7                    7                   7                 5       1
##  4    11                   23                   7                 5       1
##  5     3                   23                   7                 6       1
##  6    10                   22                   7                 6       1
##  7    20                   23                   7                 6       1
##  8    14                   19                   7                 2       1
##  9     1                   22                   7                 2       1
## 10    20                    1                   7                 2       1
## # ... with 690 more rows, and 16 more variables: Transportation expense <dbl>,
## #   Distance from Residence to Work <dbl>, Service time <dbl>, Age <dbl>,
## #   Work load Average/day  <dbl>, Hit target <dbl>, Disciplinary failure <dbl>,
## #   Education <dbl>, Son <dbl>, Social drinker <dbl>, Social smoker <dbl>,
```

```
## #   Pet <dbl>, Weight <dbl>, Height <dbl>, Body mass index <dbl>,
## #   Absenteeism time in hours <dbl>
```

Recode day of week and season Recode the columns Day.of.the.week and Seasons to the character values given above. Here are some example rows and columns.

```
absent$`Day of the week`<-recode(absent$`Day of the week`, '2'="Mon","3"="Tue","4"="Wed","5"="Thu","6"=
absent$Seasons<-recode(absent$Seasons, "1"="Summer","2"="Autumn","3"="Winter","4"="Spring")
absent
```

```
## # A tibble: 700 x 21
##       ID `Reason for absence` `Month of absence` `Day of the week` Seasons
##    <dbl>                <dbl>               <dbl> <chr>             <chr>
##  1    11                   26                   7 Tue               Summer
##  2     3                   23                   7 Wed               Summer
##  3     7                    7                   7 Thu               Summer
##  4    11                   23                   7 Thu               Summer
##  5     3                   23                   7 Fri               Summer
##  6    10                   22                   7 Fri               Summer
##  7    20                   23                   7 Fri               Summer
##  8    14                   19                   7 Mon               Summer
##  9     1                   22                   7 Mon               Summer
## 10    20                    1                   7 Mon               Summer
## # ... with 690 more rows, and 16 more variables: Transportation expense <dbl>,
## #   Distance from Residence to Work <dbl>, Service time <dbl>, Age <dbl>,
## #   Work load Average/day  <dbl>, Hit target <dbl>, Disciplinary failure <dbl>,
## #   Education <dbl>, Son <dbl>, Social drinker <dbl>, Social smoker <dbl>,
## #   Pet <dbl>, Weight <dbl>, Height <dbl>, Body mass index <dbl>,
## #   Absenteeism time in hours <dbl>
```

Proportion of absences in Day by Season Create a tibble that shows the proportion of absences in each season that occur on each weekday. Here are some of the rows in my dataset: A tibble named absence_props showing the proportion of absences that took place on a particular weekday within a season.

```
absence_props<-absent %>%
   group_by(Seasons,`Day of the week`) %>%
   count() %>%
   group_by(Seasons) %>%
   mutate(prop= n/sum(n)) %>%
   ungroup()
absence_props
```

```
## # A tibble: 20 x 4
##    Seasons `Day of the week`     n  prop
##    <chr>   <chr>             <int> <dbl>
##  1 Autumn  Fri                  34 0.178
##  2 Autumn  Mon                  42 0.220
##  3 Autumn  Thu                  34 0.178
##  4 Autumn  Tue                  42 0.220
##  5 Autumn  Wed                  39 0.204
##  6 Spring  Fri                  36 0.207
##  7 Spring  Mon                  35 0.201
##  8 Spring  Thu                  31 0.178
##  9 Spring  Tue                  36 0.207
## 10 Spring  Wed                  36 0.207
## 11 Summer  Fri                  29 0.176
## 12 Summer  Mon                  39 0.236
```

```
## 13 Summer  Thu                  29 0.176
## 14 Summer  Tue                  41 0.248
## 15 Summer  Wed                  27 0.164
## 16 Winter  Fri                  40 0.235
## 17 Winter  Mon                  39 0.229
## 18 Winter  Thu                  24 0.141
## 19 Winter  Tue                  23 0.135
## 20 Winter  Wed                  44 0.259
```

Summaries by courier For each courier, compute the following summary statistics: min, max, median, lower quartile, upper quartile, total absence time and total number of absences. Who are the most dilligent couriers? Who are the least?

Couriers 4 and 35 are the most diligent as they have only been absent once, whereas Couriers 3 and 38 have been absent for 113 and 77 times respectively.

```
statistic<- function(x){
   q_df<-as.data.frame(t(quantile(x)))
   colnames(q_df)<-c ("min","lower","median","upper","max")
   q_df$abs_time <- sum(x)
   q_df$abs_count<- length(x)
   q_df
}
courier_summary<- absent %>% group_by(ID) %>%
   summarise(statistic(`Absenteeism time in hours`),.groups="drop")
courier_summary
```

```
## # A tibble: 36 x 8
##       ID   min lower median upper   max abs_time abs_count
##    <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>    <dbl>     <int>
##  1     1     1  3         4     8    16      121        22
##  2     2     1  6.25      8     8     8       25         4
##  3     3     1  2         3     4    32      482       112
##  4     4     0  0         0     0     0        0         1
##  5     5     2  8         8     8    16      104        14
##  6     6     8  8         8     8    16       72         8
##  7     7     2  3.5       6    10    16       30         4
##  8     8     0  0         0     0     0        0         1
##  9     9     1  2.75      8    34   120      262         8
## 10    10     1  6.25      8     8    40      186        24
## # ... with 26 more rows
```

Status Changes The following demographic variables could have changed over the three years for which the couriers were tracked: Education to Pet status (five columns). Use dplyr to investigate which of the couriers' status changed over the three years.

Since there are no rows from the code run below, the couriers status have not changed over the three years.

```
helper<-function(x){
   change <- length(unique(x)) !=1
   change
}
absent %>%
   group_by(ID) %>%
   summarise(across(.cols=c(Education:Pet,`Body mass index`),helper)) %>%
   ungroup() %>%
   rowwise() %>%
```

```
  mutate(any_change=any(c_across(-1))) %>%
  filter(any_change)
```

```
## # A tibble: 0 x 8
## # Rowwise:
## # ... with 8 variables: ID <dbl>, Education <lgl>, Son <lgl>,
## #   Social drinker <lgl>, Social smoker <lgl>, Pet <lgl>,
## #   Body mass index <lgl>, any_change <lgl>
```