R Notebook

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

```
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(stringr)

#import data
montana <- read.csv("MT_cleaned.csv", stringsAsFactors = FALSE)
montanaT <- montana

vermont <- read.csv("VT_cleaned.csv", stringsAsFactors = FALSE)
vermontT <- vermont
```

```
#understand data
dim(montana)
```

```
## [1] 825118     33
```

```
str(montana)
```

```
## 'data.frame':    825118 obs. of  33 variables:
##  $ id                   : chr  "MT-2009-00001" "MT-2009-00002" "MT-2009-00003"
"MT-2009-00004" ...
##  $ state                : chr  "MT" "MT" "MT" "MT" ...
##  $ stop_date            : chr  "2009-01-01" "2009-01-02" "2009-01-03" "2009-01-0
4" ...
##  $ stop_time            : chr  "02:10" "11:34" "11:36" "10:33" ...
##  $ location_raw         : chr  "CASCADE" "MISSOULA" "MISSOULA" "MISSOULA" ...
##  $ county_name          : chr  "Cascade County" "Missoula County" "Missoula Coun
ty" "Missoula County" ...
##  $ county_fips          : int  30013 30063 30063 30063 30063 30081 30111 30063
30111 30111 ...
##  $ fine_grained_location: chr  "US 89 N MM10 (SB)" "HWY 93 SO AND ANNS LANE S/B"
"P007 HWY 93 MM 77 N/B" "P007 HWY 93 MM 81 S/B" ...
##  $ police_department    : logi  NA NA NA NA NA NA ...
##  $ driver_gender        : chr  "F" "M" "M" "F" ...
##  $ driver_age_raw       : num  16 19 17 17 31 20 30 34 21 18 ...
##  $ driver_age           : num  16 19 17 17 31 20 30 34 21 18 ...
##  $ driver_race_raw      : chr  "White" "White" "White" "" ...
##  $ driver_race          : chr  "White" "White" "White" "" ...
##  $ violation_raw        : chr  "240 - INSURANCE,150 - HIT AND RUN,245 - OTHER NO
N-HAZARDOUS" "EXPIRED TAG ( 4 MONTHS OR LESS ),SEATBELT ( DRIVER ),FAULTY EQUIPMENT
" "SPEED" "SPEED" ...
##  $ violation            : chr  "Other,Paperwork,Safe movement" "Other (non-mapp
ed),Seat belt" "Speeding" "Speeding" ...
##  $ search_conducted     : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ search_type_raw      : chr  "" "" "" "" ...
##  $ search_type          : chr  "" "" "" "" ...
##  $ contraband_found     : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ stop_outcome         : chr  "Citation" "Arrest" "Arrest" "Arrest" ...
##  $ is_arrested          : logi  FALSE TRUE TRUE TRUE TRUE TRUE ...
##  $ lat                  : num  47.6 46.8 46.7 46.7 46.7 ...
##  $ lon                  : num  -112 -114 -114 -114 -114 ...
##  $ ethnicity            : chr  "N" "N" "N" "" ...
##  $ city                 : chr  "" "" "" "" ...
##  $ out_of_state         : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ vehicle_year         : chr  "1994" "1996" "1999" "2002" ...
##  $ vehicle_make         : chr  "FORD" "GMC" "GMC" "HOND" ...
##  $ vehicle_model        : chr  "EXPLORER" "TK" "YUKON" "CR-V" ...
##  $ vehicle_style        : chr  "SPORT UTILITY" "TRUCK" "SPORT UTILITY" "SPORT UT
ILITY" ...
##  $ search_reason        : chr  "" "" "" "" ...
##  $ stop_outcome_raw     : chr  "TRAFFIC CITATION,WARNING" "INFFRACTION ARREST,WA
RNING" "INFFRACTION ARREST" "INFFRACTION ARREST" ...
```

```
head(montana)
```

	▶
1	
2	
3	
4	
5	
6	

6 rows | 1-1 of 34 columns

```
colnames(montana)
```

```
##  [1] "id"                      "state"
##  [3] "stop_date"               "stop_time"
##  [5] "location_raw"            "county_name"
##  [7] "county_fips"             "fine_grained_location"
##  [9] "police_department"       "driver_gender"
## [11] "driver_age_raw"          "driver_age"
## [13] "driver_race_raw"         "driver_race"
## [15] "violation_raw"           "violation"
## [17] "search_conducted"        "search_type_raw"
## [19] "search_type"             "contraband_found"
## [21] "stop_outcome"            "is_arrested"
## [23] "lat"                     "lon"
## [25] "ethnicity"               "city"
## [27] "out_of_state"            "vehicle_year"
## [29] "vehicle_make"            "vehicle_model"
## [31] "vehicle_style"           "search_reason"
## [33] "stop_outcome_raw"
```

```
# proportion of male drivers stop in MT
prop_m_stop = sum(montana$driver_gender=='M')/ dim(montana)[1]
print(prop_m_stop, digits = 10)
```

```
## [1] 0.6749749733
```

```
# arresting comparison between non_MT plate and MT plate
m <- subset(montana, out_of_state=='TRUE' & montana$is_arrested=='TRUE')
n <- subset(montana, out_of_state=='FALSE' & montana$is_arrested=='TRUE')
non_MT_arrt = dim(m)[1] / dim(n)[1]
print(non_MT_arrt, digits = 10)
```

```
## [1] 0.3993437244
```

```
# chi test for non_MT and MT arresting
chisq.test(table(montana$out_of_state=='TRUE' & montana$is_arrested=='TRUE',
                 montana$out_of_state=='FALSE' & montana$is_arrested=='TRUE'))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(montana$out_of_state == "TRUE" & montana$is_arrested ==     "TRUE",
montana$out_of_state == "FALSE" & montana$is_arrested ==     "TRUE")
## X-squared = 72.425, df = 1, p-value < 2.2e-16
```

```
# proportion of speeding
prop_speeding <- sum(montana$violation=='Speeding')/dim(montana)[1]
print(prop_speeding, digits = 10)
```

```
## [1] 0.4084021922
```

```
# proportion of DUI in VT
prop_DUI_vt <- sum(vermont$violation %in% c('DUI'))/dim(vermont)[1]
print(prop_DUI_vt, digits = 10)
```

```
## [1] 0.002643980444
```

```
# linear regression model between year and average_manufacture_vehicle
# amv stands for average_manufacture_vehicle
# extract year from date
montana$year_stop <- format(as.Date(montana$stop_date), format="%Y")
table(montana$year_stop)
```

```
##
##   2009   2010   2011   2012   2013   2014   2015   2016
##  18434 124285 122839 117487 114283 109747 115935 102097
```

```
montana$year_cars <- as.numeric((montana$vehicle_year))
```

```
## Warning: NAs introduced by coercion
```

```
amv_09 <-
  round(mean(montana$year_cars[which(montana$year_stop=='2009')],na.rm = TRUE),0)
amv_10 <-
  round(mean(montana$year_cars[which(montana$year_stop=='2010')],na.rm = TRUE),0)
amv_11 <-
  round(mean(montana$year_cars[which(montana$year_stop=='2011')],na.rm = TRUE),0)
amv_12 <-
  round(mean(montana$year_cars[which(montana$year_stop=='2012')],na.rm = TRUE),0)
amv_13 <-
  round(mean(montana$year_cars[which(montana$year_stop=='2013')],na.rm = TRUE),0)
amv_14 <-
  round(mean(montana$year_cars[which(montana$year_stop=='2014')],na.rm = TRUE),0)
amv_15 <-
  round(mean(montana$year_cars[which(montana$year_stop=='2015')],na.rm = TRUE),0)
amv_16 <-
  round(mean(montana$year_cars[which(montana$year_stop=='2016')],na.rm = TRUE),0)

Year <-c('2009','2010','2011','2012','2013','2014','2015','2016')
average_manufacture_vehicle <- c(amv_09, amv_10, amv_11, amv_12, amv_13, amv_14, am
v_15, amv_16)
dataT <- data.frame(Year, average_manufacture_vehicle)
View(dataT)

mod1 <- lm(Year ~ average_manufacture_vehicle)
summary(mod1)
```

```
##
## Call:
## lm(formula = Year ~ average_manufacture_vehicle)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5283 -0.2972 -0.0283  0.2453  0.6038
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -860.26415  195.19599  -4.407  0.00453 **
## average_manufacture_vehicle    1.43396    0.09743  14.717 6.18e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4344 on 6 degrees of freedom
## Multiple R-squared:  0.973,  Adjusted R-squared:  0.9686
## F-statistic: 216.6 on 1 and 6 DF,  p-value: 6.183e-06
```

```
# make prediction with year as variable
avm_20 <- (2020 + 860.26415)/1.43396
print(avm_20, digits = 10)
```

```
## [1] 2008.608434
```

```r
# import the combined data by operate cmd
# understand the combined data
data_comb <- read.csv("MT_VT_combine.csv", stringsAsFactors = FALSE)
dim(data_comb)
```

```
## [1] 1108404      33
```

```r
head(data_comb)
```

1
2
3
4
5
6

6 rows | 1-1 of 34 columns

```r
View(data_comb)[1:20]
```

```
## NULL
```

```r
str(data_comb)
```

```
## 'data.frame':    1108404 obs. of  33 variables:
##  $ id                   : chr  "MT-2009-00001" "MT-2009-00002" "MT-2009-00003"
"MT-2009-00004" ...
##  $ state                : chr  "MT" "MT" "MT" "MT" ...
##  $ stop_date            : chr  "2009-01-01" "2009-01-02" "2009-01-03" "2009-01-0
4" ...
##  $ stop_time            : chr  "02:10" "11:34" "11:36" "10:33" ...
##  $ location_raw         : chr  "CASCADE" "MISSOULA" "MISSOULA" "MISSOULA" ...
##  $ county_name          : chr  "Cascade County" "Missoula County" "Missoula Coun
ty" "Missoula County" ...
##  $ county_fips          : chr  "30013" "30063" "30063" "30063" ...
##  $ fine_grained_location: chr  "US 89 N MM10 (SB)" "HWY 93 SO AND ANNS LANE S/B"
"P007 HWY 93 MM 77 N/B" "P007 HWY 93 MM 81 S/B" ...
##  $ police_department    : chr  "" "" "" "" ...
##  $ driver_gender        : chr  "F" "M" "M" "F" ...
##  $ driver_age_raw       : chr  "16.0" "19.0" "17.0" "17.0" ...
##  $ driver_age           : chr  "16.0" "19.0" "17.0" "17.0" ...
##  $ driver_race_raw      : chr  "White" "White" "White" "" ...
##  $ driver_race          : chr  "White" "White" "White" "" ...
##  $ violation_raw        : chr  "240 - INSURANCE,150 - HIT AND RUN,245 - OTHER NO
N-HAZARDOUS" "EXPIRED TAG ( 4 MONTHS OR LESS ),SEATBELT ( DRIVER ),FAULTY EQUIPMENT
" "SPEED" "SPEED" ...
##  $ violation            : chr  "Other,Paperwork,Safe movement" "Other (non-mapp
ed),Seat belt" "Speeding" "Speeding" ...
##  $ search_conducted     : chr  "FALSE" "FALSE" "FALSE" "FALSE" ...
##  $ search_type_raw      : chr  "" "" "" "" ...
##  $ search_type          : chr  "" "" "" "" ...
##  $ contraband_found     : chr  "FALSE" "FALSE" "FALSE" "FALSE" ...
##  $ stop_outcome         : chr  "Citation" "Arrest" "Arrest" "Arrest" ...
##  $ is_arrested          : chr  "FALSE" "TRUE" "TRUE" "TRUE" ...
##  $ lat                  : chr  "47.5727383333333" "46.761225" "46.6946833333333
" "46.7273883333333" ...
##  $ lon                  : num  -112 -114 -114 -114 -114 ...
##  $ ethnicity            : chr  "N" "N" "N" "" ...
##  $ city                 : chr  "" "" "" "" ...
##  $ out_of_state         : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ vehicle_year         : chr  "1994" "1996" "1999" "2002" ...
##  $ vehicle_make         : chr  "FORD" "GMC" "GMC" "HOND" ...
##  $ vehicle_model        : chr  "EXPLORER" "TK" "YUKON" "CR-V" ...
##  $ vehicle_style        : chr  "SPORT UTILITY" "TRUCK" "SPORT UTILITY" "SPORT UT
ILITY" ...
##  $ search_reason        : chr  "" "" "" "" ...
##  $ stop_outcome_raw     : chr  "TRAFFIC CITATION,WARNING" "INFRACTION ARREST,WA
RNING" "INFRACTION ARREST" "INFRACTION ARREST" ...
```

```r
# extract hours from the combined data
Split <- strsplit(as.character(data_comb$stop_time), ":", fixed = TRUE)
data_comb$stop_hs <- sapply(Split, "[", 1)
table(data_comb$stop_hs)
```

```
##
##         00         01         02         03         04         05         06
##      25490      16856       8399       1425        547       1710       8561
##         07         08         09         10         11         12         13
##      41550      62488      62233      61946      51008      44024      59281
##         14         15         16         17         18         19         20
##      82129      95891      86886      81437      82430      57980      47244
##         21         22         23  stop_time
##      45891      44387      38599          1
```

```
Split <- strsplit(as.character(montana$stop_time), ":", fixed = TRUE)
montana$stop_hs <- sapply(Split, "[", 1)
sort(table(montana$stop_hs))
```

```
##
##     04     03     05     06     02     01     00     23     22     07     21     12
##    229    681   1092   5473   6202  10405  14923  25702  31843  32936  34275  34694
##     20     11     19     09     13     08     10     18     17     14     16     15
## 36281  40166  42050  45386  46078  47336  47519  56060  57549  64637  67883  75707
```

```
diff_stop_num=75707 -229
print(diff_stop_num, digits = 10)
```

```
## [1] 75478
```

```
#predict county area with longitude and latitude
data5 <- group_by(montana, county_name)
head(data5)
```

id <chr> ▶
MT-2009-00001
MT-2009-00002
MT-2009-00003
MT-2009-00004
MT-2009-00005
MT-2009-00006

6 rows | 1-1 of 36 columns

```
new_f <-summarise(data5,
                  count=n(),
                  lat_sd=sd(lat, na.rm = TRUE),
                  lon_sd=sd(lon, na.rm = TRUE))
View(new_f)
new_f$size_sqkm <- 3.1415926535 * 2 *new_f$lat_sd * 2 * new_f$lon_sd *10
print(max(new_f$size_sqkm), digits = 10)
```

```
## [1] 3194.220151
```

Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.