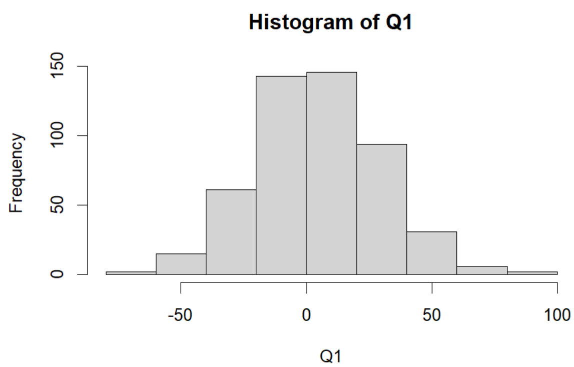
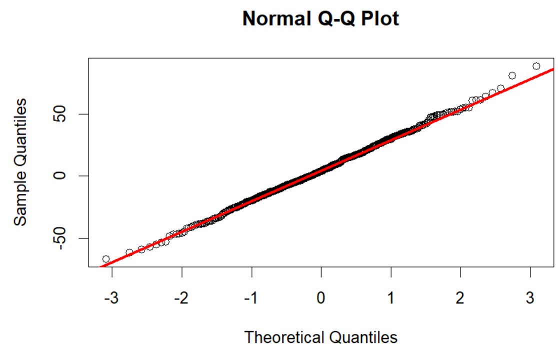
STAT 40001/STAT 50001 Statistical Computing Fall 2024

Lab-8

1) Generate 500 random numbers from a normal distribution with mean 4 and variance 25.
   a) How many observations are greater than 4?
   b) Draw a histogram of the data that you generated
   c) Display the quantile-quantile plot

```
> Q1 = rnorm(500,mean=4,sd=25)
> sum(Q1>4)
[1] 247
> #length(Q1[Q1>4])
> length(Q1)
[1] 500
> hist(Q1)
> qqnorm(Q1)
> qqline(Q1, col = "red", lwd = 3)
```





2) All registered elevators in New York City are provided in the link below
   https://www.kaggle.com/new-york-city/nyc-elevators/discussion/39528

   For your convenience the data are attached with this Lab-8 (NYC data)
   a) Import the data in R.
   b) Not all elevators are active. How many are currently active?
   c) How many elevators are active in Manhattan borough?

```
> Q2 <- read.csv("C:\\Users\\PNW_checkout\\Downloads\\vaishak\\PNW_COURSE-WORK\\FALL24\\STATISTIC
AL COMPUTING\\Assignment\\ASSIGNMENT 8\\archive\\nyc-elevators.csv")
> dim(Q2)
[1] 76088    29
> variable.names(Q2)
 [1] "DV_DEVICE_NUMBER"             "Device.Status"
 [3] "DV_DEVICE_STATUS_DESCRIPTION" "BIN"
 [5] "TAX_BLOCK"                    "TAX_LOT"
 [7] "HOUSE_NUMBER"                 "STREET_NAME"
 [9] "ZIP_CODE"                     "Borough"
[11] "Device.Type"                  "DV_LASTPER_INSP_DATE"
[13] "DV_LASTPER_INSP_DISP"         "DV_APPROVAL_DATE"
[15] "DV_MANUFACTURER"              "DV_TRAVEL_DISTANCE"
[17] "DV_SPEED_FPM"                 "DV_CAPACITY_LBS"
[19] "DV_CAR_BUFFER_TYPE"           "DV_GOVERNOR_TYPE"
[21] "DV_MACHINE_TYPE"              "DV_SAFETY_TYPE"
[23] "DV_MODE_OPERATION"            "DV_STATUS_DATE"
[25] "DV_FLOOR_FROM"                "DV_FLOOR_TO"
[27] "X"                            "LATITUDE"
[29] "LONGITUDE"
```

```
> table(Q2$DV_DEVICE_STATUS_DESCRIPTION)

    ACTIVE DISMANTLED  NO JURIST      NYCHA WK IN PROG
     66885       1506       1983       3365       2349
> Q21 <- subset(Q2, DV_DEVICE_STATUS_DESCRIPTION == "ACTIVE")
> dim(Q21)
[1] 66885    29
> Q22 <- subset(Q2, Borough == "Manhattan")
> table(Q22$DV_DEVICE_STATUS_DESCRIPTION)

    ACTIVE DISMANTLED  NO JURIST      NYCHA WK IN PROG
     39379        965       1105        951       1196
```

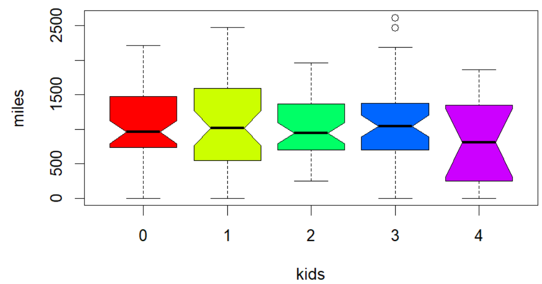
3) The data "vacation" provided in the link below describe a sample of 200 Chicago households regarding their vacation. The data includes the following variables

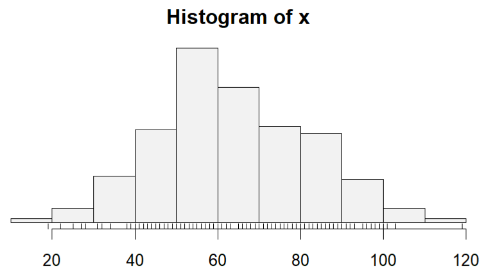
```
 1. miles        miles traveled per year
 2. income       annual income in $1000's
 3. age          average age of adult members of household
 4. kids         number of children in household
```

http://www.principlesofeconometrics.com/poe4/poe4stata.htm

a) Import the data in R (Note the format of the data)
b) Display the miles distribution based on the number of kids by drawing parallel box-plot
c) Draw histogram along with boxplot of the income data. (You will need to use UsingR packages and simple.hist.and.boxplot(your data)

```
> install.packages("haven")
> library(haven)
> url = "http://www.principlesofeconometrics.com/poe4/data/stata/vacation.dta"
> Q3 = read_dta(url)
> head(Q3)
# A tibble: 6 × 4
  miles income   age  kids
  <dbl>  <dbl> <dbl> <dbl>
1   902     41    26     0
2   491     31    38     3
3  1841     87    40     2
4   406     54    48     4
5     0     77    43     4
6  1899     70    55     2
> dim(Q3)
[1] 200   4
> str(Q3)
tibble [200 × 4] (S3: tbl_df/tbl/data.frame)
 $ miles : num [1:200] 902 491 1841 406 0 ...
  ..- attr(*, "label")= chr "miles traveled per year"
  ..- attr(*, "format.stata")= chr "%8.0g"
 $ income: num [1:200] 41 31 87 54 77 70 43 40 69 38 ...
  ..- attr(*, "label")= chr "annual income ($1000)"
  ..- attr(*, "format.stata")= chr "%8.0g"
 $ age   : num [1:200] 26 38 40 48 43 55 24 39 55 29 ...
  ..- attr(*, "label")= chr "average age of adult members of household"
  ..- attr(*, "format.stata")= chr "%8.0g"
 $ kids  : num [1:200] 0 3 2 4 4 2 0 0 2 2 ...
  ..- attr(*, "label")= chr "number of children in household"
  ..- attr(*, "format.stata")= chr "%8.0g"
> attach(Q3)
> boxplot(miles~kids, col = rainbow(5), notch = T)$out
[1] 2609 2464
> install.packages("UsingR")
> library(UsingR)
> simple.hist.and.boxplot(income)
```





4) The following are the head circumferences (centimeters) at birth of 15 infants in a local hospital

33.38, 32.15, 33.99, 34.10, 33.97, 34.34, 33.95, 33.85, 34.23, 32.73, 33.46, 34.13, 34.45, 34.19 , 34.05

Construct a 90% confidence interval for the head circumferences (centimeters) at birth of all infants born at the local hospital.

```
> Q4 = c(33.38,32.15,33.99,34.10,33.97,34.34,33.95,33.85,34.23,32.73,33.46,34.13,34.45,34.19,34.0
5)
> t.test(Q4, conf.level = 0.9)$conf.int
[1] 33.51136 34.08464
attr(,"conf.level")
[1] 0.9
```

5) The National Hurricane Center (NHC) of the National Oceanic and Atmospheric Administration (NOAA) tries to predict the path each hurricane will take. But hurricanes tend to wander around aimlessly and are pushed by fronts and other weather phenomena in their area, so they are notoriously difficult to predict. Even relatively small changes in a hurricane's track can make big differences in the damage it causes. The link below give the mean error in nautical miles of the NHC's , 24, 48 and 72-hour predictions of Atlantic hurricanes for 1970-2017. NOAA refers to these errors as the Forecast error or the Prediction error and reports annual results.

https://dasl.datadescription.com/datafile/tracking-hurricanes-2016/

a) Import the data in R
b) Display the 24-, 48- and 72-hours errors creating appropriate graph.
c) Construct 90% confidence interval for 72 hours prediction errors
d) Construct a 95% confidence interval for the difference in the prediction error at 48 hours and 72 hours.

```
> Q5 <- read.table("C:\\Users\\PNW_checkout\\Downloads\\vaishak\\PNW_COURSE-WORK\\FALL24\\STATIST
ICAL COMPUTING\\Assignment\\ASSIGNMENT 8\\tracking-hurricanes-2016.txt", header = TRUE)
> head(Q5)
  Year Error_24h Error_48h Error_72h
1 1970      84.3     185.8     253.8
2 1971     112.4     242.0     381.9
3 1972     142.3     390.6     689.2
```

```
4 1973      116.7      246.2      363.2
5 1974       97.1      206.5      348.3
6 1975      117.0      256.9      402.1
>
> boxplot(Q5$Error_24h, Q5$Error_48h, Q5$Error_72h,
+         names = c("24h", "48h", "72h"))
> t.test(Q5$Error_72h, conf.level = 0.9)$conf.int
[1] 230.5047 294.0230
attr(,"conf.level")
[1] 0.9
> t.test(Q5$Error_48h, Q5$Error_72h, paired = TRUE)

	Paired t-test

data:  Q5$Error_48h and Q5$Error_72h
t = -11.028, df = 46, p-value = 1.654e-14
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -106.66883  -73.73968
sample estimates:
mean difference
      -90.20426
```