HW1_Videtti

```
#HOMEWORK 1 by Nick Videtti: I produced the work below with only help from
the textbook and asynchronous material.




#Example 1: Using the material from this chapter and possibly other
information that you look up, write a brief definition of these terms in your
own words: mean, median, mode, variance, standard deviation, histogram,
normal distribution, and Poisson distribution.

#mean - the sum of values divided by the number of values, also known as the
average

#median - value of the middle observation when all observations are listed in
order (average of two middle observations if there are an even number of
observations)

#mode - most frequent value that occurs

#variance - average squared deviation from the mean

#standard deviation - square root of the average squared deviation from the
mean, or square root of variance

#histogram - frequency plot that visualizes the distribution of data using
bins that cover a range of values

#normal distribution - "bell curve" distribution where data are symmetrically
distributed around the mean, with most values occurring at or near the mean

#Poisson distribution - distribution that models positive values where the
shape of the distribution varies and is heavily influenced by its lambda
value
################################################################################
##############################################




#Exercise 3: Use the data() function to get a list of the data sets that are
included with the basic installation of R: just type “data()” at the command
line and press enter. Choose a data set from the list that contains at least
one numeric variable—for example, the Bio- chemical Oxygen Demand (BOD) data
```

```
set. Use the summary() command to summarize the variables in the data set you
selected—for example, summary(BOD). Write a brief description of the mean and
median of each numeric variable in the data set. Make sure you define what a
“mean” and a “median” are, that is, the technical definition and practical
meaning of each of these quantities.

data()

?WorldPhones

## starting httpd help server ... done

WorldPhones

##      N.Amer Europe Asia S.Amer Oceania Africa Mid.Amer
## 1951  45939  21574 2876   1815    1646     89      555
## 1956  60423  29990 4708   2568    2366   1411      733
## 1957  64721  32510 5230   2695    2526   1546      773
## 1958  68484  35218 6662   2845    2691   1663      836
## 1959  71799  37598 6856   3000    2868   1769      911
## 1960  76036  40341 8220   3145    3054   1905     1008
## 1961  79831  43173 9053   3338    3224   2005     1076

summary(WorldPhones)

##      N.Amer          Europe           Asia          S.Amer        Oceania
##  Min.   :45939   Min.   :21574   Min.   :2876   Min.   :1815   Min.
:1646
##  1st Qu.:62572   1st Qu.:31250   1st Qu.:4969   1st Qu.:2632   1st
Qu.:2446
##  Median :68484   Median :35218   Median :6662   Median :2845   Median
:2691
##  Mean   :66748   Mean   :34343   Mean   :6229   Mean   :2772   Mean
:2625
##  3rd Qu.:73918   3rd Qu.:38970   3rd Qu.:7538   3rd Qu.:3072   3rd
Qu.:2961
##  Max.   :79831   Max.   :43173   Max.   :9053   Max.   :3338   Max.
:3224
##      Africa        Mid.Amer
##  Min.   :  89   Min.   : 555.0
##  1st Qu.:1478   1st Qu.: 753.0
##  Median :1663   Median : 836.0
##  Mean   :1484   Mean   : 841.7
##  3rd Qu.:1837   3rd Qu.: 959.5
##  Max.   :2005   Max.   :1076.0

#We have data here that show the number of telephones in 8 different regions
in the years 1951, 1956, 1957, 1958, 1959, 1960, and 1961. All 8 variables
are numeric, and they are simply the number of telephones in each region.

#As we discussed in Exercise 1, mean is the sum of values divided by the
```

```
number of values, also known as the average, and median is the value of the
middle observation when all observations are listed in order (average of two
middle observations if there are an even number of observations).

#This tells us that the mean of each of these variables is the average number
of telephones in each region over the 7 different years that we have data
for. Also, the median of each of these variables is the middle value (when
values are listed in order) for the number of telephones in each region over
the 7 different years we have data for. For example, the mean of the Europe
variable is the average number of telephones that were in Europe between
1951, 1956, 1957, 1958, 1959, 1960, and 1961, and the median of Europe is the
middle value (if all values are listed in order) of the number of telephones
that were in Europe between 1951, 1956, 1957, 1958, 1959, 1960, and 1961.
##############################################################################
###################################
```

```
#Exercise 4: As in the previous exercise, use the data() function to get a
list of the data sets that are included with the basic installation of R.
Choose a data set that includes just one variable, for example, the LakeHuron
data set (levels of Lake Huron in the years 1875 through 1972). Use the
hist() command to create a histogram of the variable—for example,
hist(LakeHuron). Describe the shape of the histogram in words. Which of the
distribution types do you think these data fit most closely (e.g., normal,
Poisson). Speculate on why your selected data may fit that distribution.
```

```
data()
```

```
airmiles
```

```
## Time Series:
## Start = 1937
## End = 1960
## Frequency = 1
##  [1]   412   480   683  1052  1385  1418  1634  2178  3362  5948  6109
5981
## [13]  6753  8003 10566 12528 14760 16769 19819 22362 25340 25343 29269
30514
```

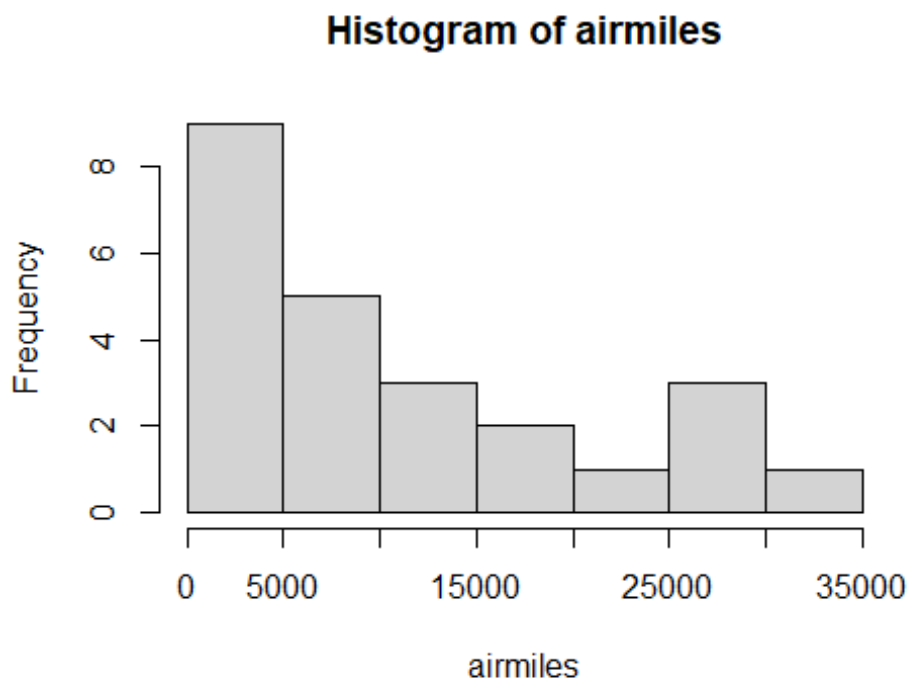
```
data.frame(airmiles)
```

```
##    airmiles
## 1       412
## 2       480
## 3       683
## 4      1052
## 5      1385
## 6      1418
```

```
## 7      1634
## 8      2178
## 9      3362
## 10     5948
## 11     6109
## 12     5981
## 13     6753
## 14     8003
## 15    10566
## 16    12528
## 17    14760
## 18    16769
## 19    19819
## 20    22362
## 21    25340
## 22    25343
## 23    29269
## 24    30514
```

```
?airmiles
```

```
hist(airmiles)
```




```
#These data appear to have a right-skewed distribution. The data also appear
to be distributed only across positive values, and this distribution is
certainly not anywhere close to symmetrical. For these reasons, these data
seem to most closely fit Poisson distribution.
```