Task 1 Data Transformation

```
1. who1 <- who %>%
     pivot_longer(
       cols = new_sp_m014:newrel_f65,
       names_to = "key",
       values_to = "cases",
       values_drop_na = TRUE
     )

2. library("dplyr")# install.packages("dplyr") for data manupulation
   who2 <- who1 %>%
   mutate(
       key = stringr::str_replace(key, "newrel", "new_rel")
     )

3. who3<-who2 %>% separate(key, c("new", "type", "sexage"), sep="_")# the purpose
   of using %>%  is to connect multiple verb actions together into a pipeline
   who3<-who3 %>%
    select(-new, -iso2, -iso3) # dropped the new column because it's constant and
   also dropped iso2 and iso3 because they are redundant

4. who4 <- who3 %>%
     separate(sexage, c("sex", "age"), sep = 1)

5. head(who4,5) #print the first 5 rows## # A tibble: 5 x 6
   ##   country      year type  sex   age   cases
   ##   <chr>       <int> <chr> <chr> <chr> <int>
   ## 1 Afghanistan  1997 sp    m     014       0
   ## 2 Afghanistan  1997 sp    m     1524     10
   ## 3 Afghanistan  1997 sp    m     2534      6
   ## 4 Afghanistan  1997 sp    m     3544      3
   ## 5 Afghanistan  1997 sp    m     4554      5
```

```
    tail(who4,5) #print the last 5 rows ## # A tibble: 5 x 6
##   country   year type  sex   age   cases
##   <chr>    <int> <chr> <chr> <chr> <int>
## 1 Zimbabwe  2013 rel   f     2534   4649
## 2 Zimbabwe  2013 rel   f     3544   3526
## 3 Zimbabwe  2013 rel   f     4554   1453
## 4 Zimbabwe  2013 rel   f     5564    811
## 5 Zimbabwe  2013 rel   f     65      725
```

```
6. write.csv(who4, "who4.csv", row.names = FALSE)# save who4 to csv
```

Task 2 Basic Statistics for Data Science in R

1.
```r
mean(Nile) # compute the mean
```

```r
## [1] 919.35

median(Nile) #compute the median

## [1] 893.5

mode(Nile)# compute the mode

## [1] "numeric"

var(Nile)# compute the variance

## [1] 28637.95

sd(Nile)# compute the standard deviation

## [1] 169.2275
```

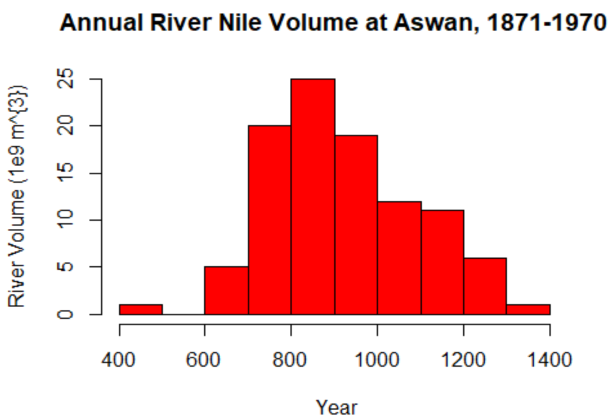
2.
```r
min(Nile)# compute the minimum
```

```r
## [1] 456

max(Nile)# compute the maximum

## [1] 1370

range(Nile)#compute the range

## [1]  456 1370
```

3.
```r
IQR(Nile)#the difference between Q1 and Q3 based on the quantile
```

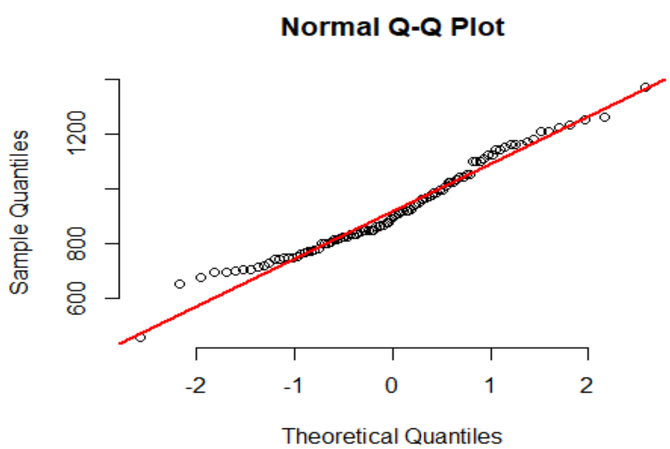
```r
## [1] 234

quantile(Nile)#returns the five number summary that shows the min,  Q1, median, Q3,
max

##     0%    25%    50%    75%   100%
##  456.0  798.5  893.5 1032.5 1370.0
```

4.
```r
hist(Nile, xlab = " Year", ylab = "River Volume (1e9 m^{3})",
     main = "Annual River Nile Volume at Aswan, 1871-1970"
     , col = "red") # the default xlab is "Time"
```



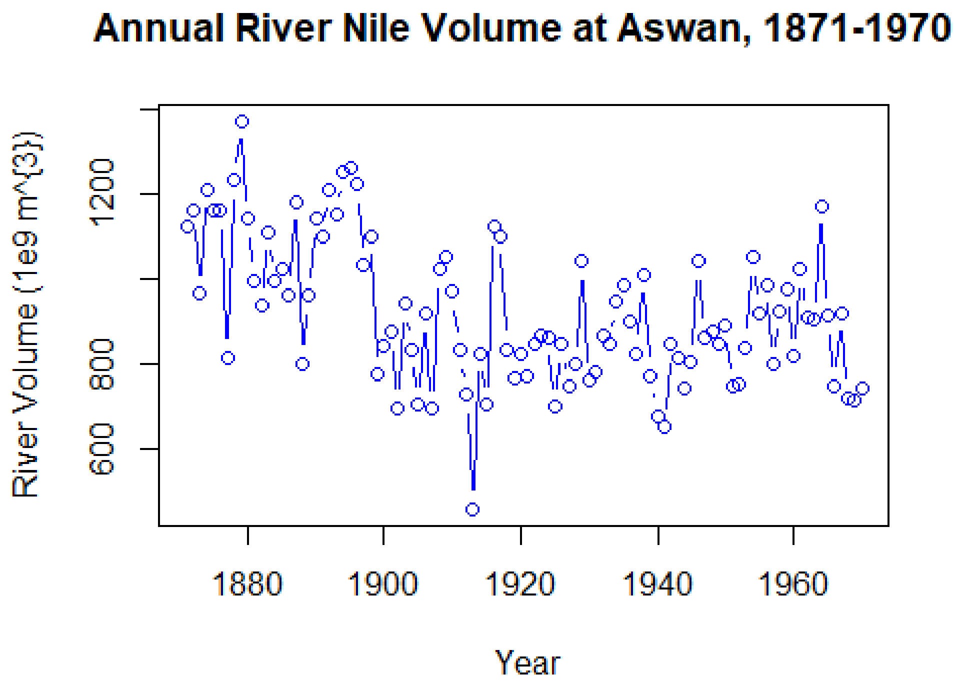
On the River Nile histogram, each of the 9 bars spans an interval of 400, beginning at 400 and ending at 1400 (Year). The tallest bar at the center of the figure covers 800 to 900 range and has a frequency of 25(river volume).it is positively skewed with a skewness of 0.3221274

5.
```r
qqnorm(Nile, pch = 1, frame = FALSE)
qqline(Nile, col = "red", lwd = 2)# adds a reference line
```



The line determines whether the values in the plot fall on a straight line.The data points near the tails don’t fall exactly along the straight line, but for the most part this sample data appears to be normally distributed.

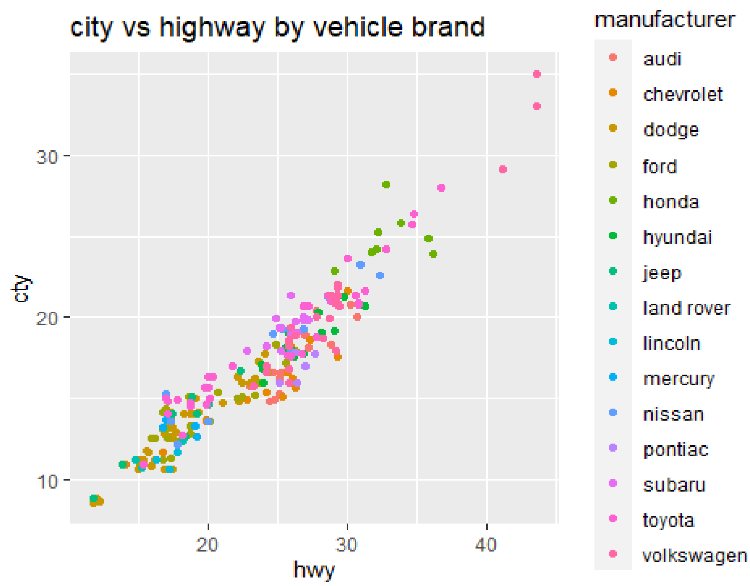
6.
```
plot(Nile, xlab = " Year", ylab = "River Volume (1e9 m^{3})",
      main = "Annual River Nile Volume at Aswan, 1871-1970",
      type = "b", col = "blue")
```



The plot() function explores the data showing the River volume and the year. The default xlab is “Time”

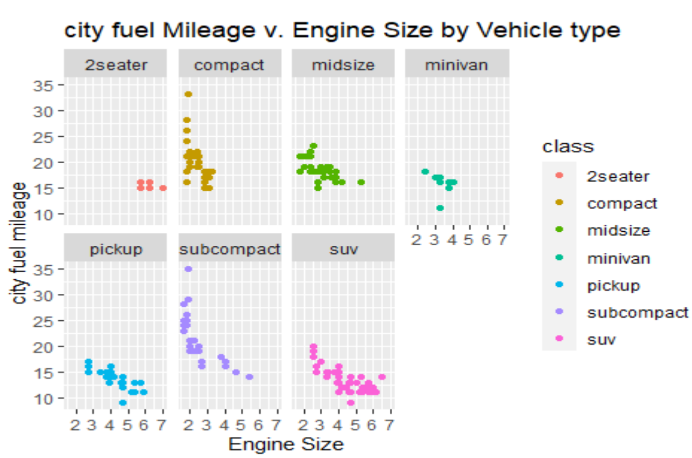
Task 3 Visualization with ggplot2

```
1. library(ggplot2)
   ggplot(data=mpg, mapping = aes(x=hwy, y=cty, color = manufacturer))+
     geom_jitter()+labs(title="city vs highway by vehicle brand",
   x="hwy",y="cty")
```



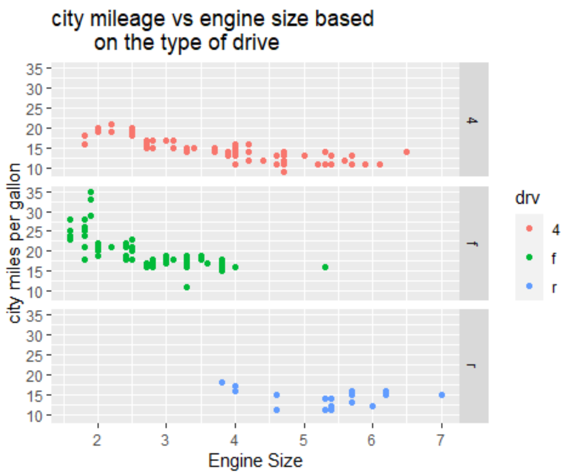
The plot shows a correlation between the manufacturer (vehicle brand), city mpg and highway mpg. Toyota and Volkswagen tends to have high mpg in both the city and highway hence they offer the best mpg in both city and in the highway.

```
2. ggplot(data = mpg) +
     geom_point(mapping = aes(x=displ, y=cty, color =class))+
     facet_wrap(~class, nrow = 2)+
     labs(title="city fuel Mileage v. Engine Size by Vehicle type", x="Engine
   Size",y="city fuel mileage")
```



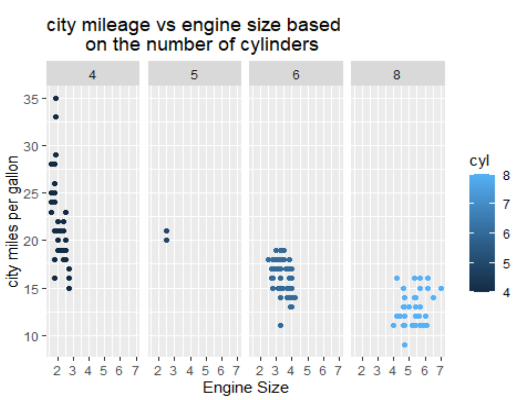
As the engine size decreases, the city fuel milleage increases and as the engine size increases, the city fuel milleage decreases. pickup and Suv have the lowest mpg in the city. The plot shows that Suv and Pickup have low city mileage and high engine displacement.That means that high displacement engines use more fuel.

3.
```
#creating facets vertically on drive (cty  vs displ)
ggplot(data = mpg) +
  geom_point(mapping = aes(x=displ, y=cty, color =drv))+
  facet_grid(drv ~ .)+
  labs(title="city mileage vs engine size based
       on the type of drive",
       x="Engine Size",y="city miles per gallon")
```



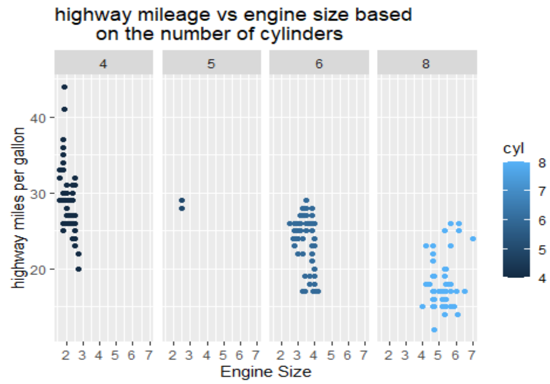
Above plot shows the relationship between city miles per gallon and engine displacement. The higher an engine's displacement, the more fuel it will consume, thus the lower mileage per gallon. The front wheel drive category streches out from 35 to 11 mpg. The front wheel drives outperform rear wheel and 4wd(four wheel drive). Rear wheel mpg is between 10-17. All four wheels'mpg are below 25. Subcompact cars have high mpg which is similar with the front wheel drive. It is safe to say that most subcompact cars are front wheel drive.

```
#creating facets horizontally on cylinders (cty  vs displ)
ggplot(data = mpg) +
  geom_point(mapping = aes(x=displ, y=cty, color =cyl))+
  facet_grid(. ~cyl)+
  labs(title="city mileage vs engine size based
       on the number of cylinders",
       x="Engine Size",y="city miles per gallon")
```



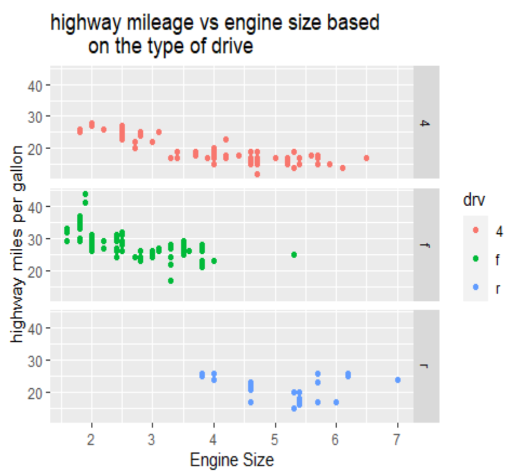
Vehicles with more cylinders consumes more fuel. The plot suggest that Vehicles fitted with 4 cylinder engines have a high mpg of 35 in the city. Subcompact car choose to fit the 4 cylinder engine and more energy efficient. These features determine they consume less fuel which results to the best mpg performance in terms of the city miles.

```r
#creating facets horizontally on cylinders (hwy  vs displ)
ggplot(data = mpg) +
  geom_point(mapping = aes(x=displ, y=hwy, color =cyl))+
  facet_grid(. ~cyl)+
  labs(title="highway mileage vs engine size based
       on the number of cylinders",
       x="Engine Size",y="highway miles per gallon")
```



Vehicles with more cylinders consumes more fuel. The plot suggest that Vehicles fitted with 4 cylinder engines have a high mpg of 44 in the highway.

```r
#creating facets vertically on drive (hwy  vs displ)
ggplot(data = mpg) +
  geom_point(mapping = aes(x=displ, y=hwy, color =drv))+
  facet_grid(drv ~ .)+
  labs(title="highway mileage vs engine size based
       on the type of drive",
       x="Engine Size",y="highway miles per gallon")
```



Above plot shows the relationship between highway miles per gallon and engine displacement. The front wheel drive category streches out from 44 to 14 mpg. The front wheel drives outperform rear wheel and 4wd(four wheel drive). All four wheels' and rear wheels mpg are below 30.

I would choose the Subcompact car, they have high mpg with relative high displacement engine and they drive mostly on the highway.