S. Afreen

21BEC1017

PROBABILITY AND STATISTICS

WORLD BANK DATA

SOURCE OF THE DATASET : https://data.worldbank.org/

NUMBER OF DATAPOINTS : 5

NUMBER OF VARIABLES : 9

VARIABLES :

country, iso2c, iso3c, year, elec, cab, edb, cpi, rate

elec-Access to electricity (% of population)

cab-Current account balance (% of GDP)

edb-Ease of doing business

cpi-Inflation, consumer prices (annual %)

rate-Interest rate spread (lending rate minus deposit rate, %)

 # Installing packages and getting basic values related to probability and statistics

```
install.packages("WDI")

library(tidyverse)

library(WDI)

data=WDI(indicator = c(elec="EG.ELC.ACCS.ZS", # access to electricity
                       cab="BN.CAB.XOKA.GD.ZS", # current account balance
                       edb="IC.BUS.DFRN.XQ", # ease of doing business
                       cpi="FP.CPI.TOTL.ZG", # CPI
                       rate="FR.INR.LNDP"), # interest rate spread
    start = 1960, end = 2020) %>% as_tibble()

summary(data)#provides descriptive statistics

str(data)#structure of data set

mean(data$elec,na.rm=TRUE)
```

median(data$cab,na.rm=TRUE)

mode(data$edb)

```
> data=WDI(indicator = c(elec="EG.ELC.ACCS.ZS", # access to electricity
+                        cab="BN.CAB.XOKA.GD.ZS", # current account balance
+                        edb="IC.BUS.DFRN.XQ", # ease of doing business
+                        cpi="FP.CPI.TOTL.ZG", # CPI
+                        rate="FR.INR.LNDP"), # interest rate spread
+     start = 1960, end = 2020) %>% as_tibble()
> summary(data)#provides descriptive statistics
   country             iso2c
 Length:16226       Length:16226
 Class :character   Class :character
 Mode  :character   Mode  :character

    iso3c                year          elec
 Length:16226       Min.   :1960   Min.   :  0.534
 Class :character   1st Qu.:1975   1st Qu.: 67.332
 Mode  :character   Median :1990   Median : 98.624
                    Mean   :1990   Mean   : 80.603
                    3rd Qu.:2005   3rd Qu.:100.000
                    Max.   :2020   Max.   :100.000
                                   NA's   :9131
      cab                 edb              cpi
 Min.   :-240.521   Min.   :19.98   Min.   :  -18.109
 1st Qu.:  -7.135   1st Qu.:51.94   1st Qu.:    2.353
 Median :  -2.919   Median :60.02   Median :    4.879
 Mean   :  -2.952   Mean   :60.86   Mean   :   20.563
 3rd Qu.:   0.940   3rd Qu.:71.82   3rd Qu.:    9.800
 Max.   : 311.761   Max.   :87.17   Max.   :23773.132
 NA's   :9467       NA's   :15039   NA's   :5927
      rate
 Min.   :-30827.437
 1st Qu.:     3.943
 Median :     6.254
 Mean   :     1.031

 3rd Qu.:     8.379
 Max.   :  1820.451
 NA's   :11522
> str(data)#structure of data set
tibble [16,226 × 9] (S3: tbl_df/tbl/data.frame)
 $ country: chr [1:16226] "Afghanistan" "Afghanistan" "Afghanistan" "Af
ghanistan" ...
 $ iso2c  : chr [1:16226] "AF" "AF" "AF" "AF" ...
 $ iso3c  : chr [1:16226] "AFG" "AFG" "AFG" "AFG" ...
 $ year   : int [1:16226] 1960 1961 1962 1963 1964 1965 1966 1967 1968
 1969 ...
 $ elec   : num [1:16226] NA NA NA NA NA NA NA NA NA NA ...
  ..- attr(*, "label")= chr "Access to electricity (% of population)"
 $ cab    : num [1:16226] NA NA NA NA NA NA NA NA NA NA ...
  ..- attr(*, "label")= chr "Current account balance (% of GDP)"
 $ edb    : num [1:16226] NA NA NA NA NA NA NA NA NA NA ...
  ..- attr(*, "label")= chr "Ease of doing business score (0 = lowest p
erformance to 100 = best performance)"
 $ cpi    : num [1:16226] NA NA NA NA NA NA NA NA NA NA ...
```

```
  ..- attr(*, "label")= chr "Inflation, consumer prices (annual %)"
 $ rate   : num [1:16226] NA NA NA NA NA NA NA NA NA NA ...
  ..- attr(*, "label")= chr "Interest rate spread (lending rate minus d
eposit rate, %)"
> mean(data$elec,na.rm=TRUE)
[1] 80.60263
> median(data$cab,na.rm=TRUE)
[1] -2.918621
> mode(data$edb)
[1] "numeric"
```

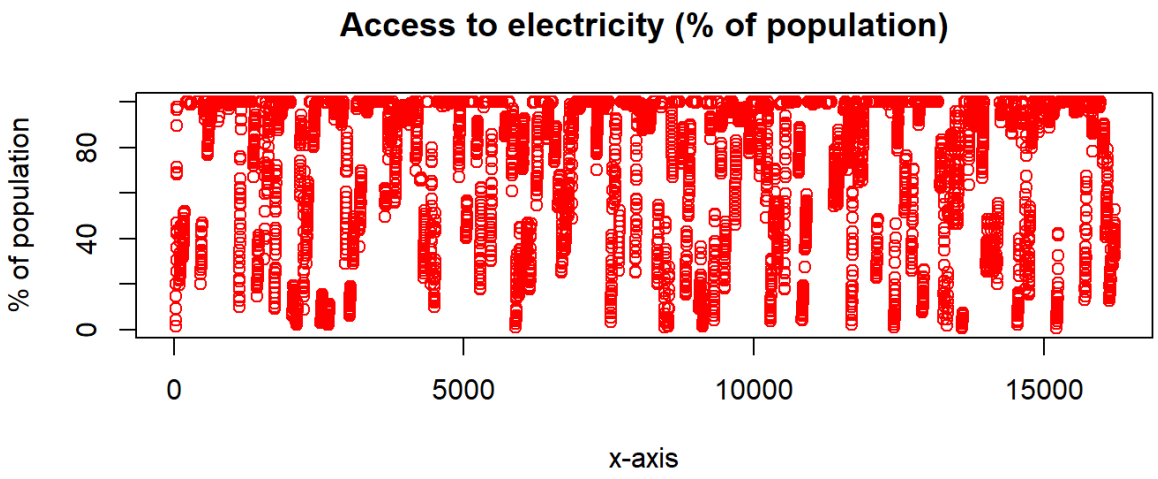
 # Extracting data, obtaining different types of tables, and graphical reperesentations

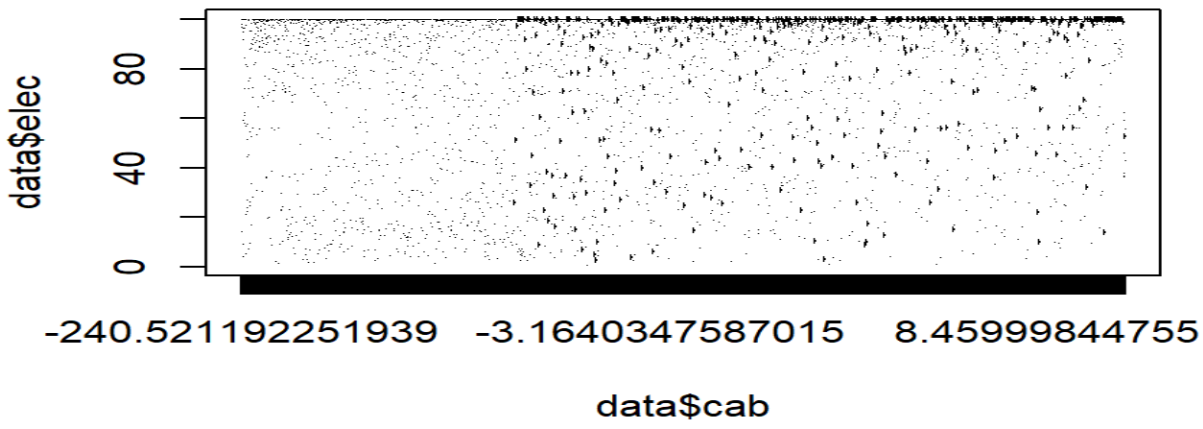
```
# Extract data related to Kuwait

kuwait=subset(data,data$elec=="Kuwait",na.rm=TRUE)

kuwait

# Extract data related to Tuvalu

tuvalu=subset(data, data$elec=="Tuvalu",na.rm=TRUE)

tuvalu

# Creating table (one-way)

table1=table(data$elec)

table1

table2=table(data$cab)

table2

# Creating table (two-way)

table3=table(data$elec, data$cab)

table3

# Graphical representation (scatter plot)

plot(data$elec,type="p",main="Access to electricity (% of
population)",xlab="x-axis",ylab="% of population",col="red")

# Graphical representation (Box plot)

boxplot(data$elec~data$cab,col=c('red','blue'))
```

```
> # Extract data related to Kuwait
> kuwait=subset(data,data$elec=="Kuwait",na.rm=TRUE)
> kuwait
# A tibble: 0 × 9
# … with 9 variables: country <chr>, iso2c <chr>, iso3c <chr>,
#   year <int>, elec <dbl>, cab <dbl>, edb <dbl>, cpi <dbl>,
#   rate <dbl>
# ℹ Use `colnames()` to see all variable names
> # Extract data related to Tuvalu
> tuvalu=subset(data, data$elec=="Tuvalu",na.rm=TRUE)
> tuvalu
# A tibble: 0 × 9
# … with 9 variables: country <chr>, iso2c <chr>, iso3c <chr>,
#   year <int>, elec <dbl>, cab <dbl>, edb <dbl>, cpi <dbl>,
#   rate <dbl>
# ℹ Use `colnames()` to see all variable names
> # Creating table (one-way)
> table1=table(data$elec)
> table1

0.533898532390594 0.643131792545319  1.02783596515656
                1                 1                 1
 1.03156232833862  1.25226926803589  1.25370573997498
                1                 1                 1
 1.27928960323334  1.42722380161285               1.5
                1                 1                 1
 1.55355584621429  1.61359095573425  1.89250123500824
                1                 1                 1
 1.89999997615814  2.01366138458252  2.07094931602478
                1                 1                 1
 2.15848255157471  2.17412757873535               2.3
                1                 1                 2
 2.33018779754639  2.42038726806641  2.46323680877686
                1                 1                 1
 2.47172713279724  2.59146237373352  2.66000008583069
                1                 1                 1

> table2=table(data$cab)
> table2

-240.521192251939 -147.997303808308 -86.9941807626357
                1                 1                 1
-70.4280506286334 -65.0289253711054  -60.880077895534
                1                 1                 1
-60.2162448692878 -59.9955509928998 -55.9089622190359
                1                 1                 1
-52.6914877540647 -52.5178130200926 -52.4853470186806
                1                 1                 1
-52.2800998128509 -49.7318371902327 -49.6472351090603
                1                 1                 1
-48.0293948585418 -47.2996244934574 -45.6874758761252
                1                 1                 1
-45.1079562259686 -44.8409951545259 -44.5332483569807
                1                 1                 1
-44.0158231604942 -43.7712346806092 -43.6608427039024
```

#Two-way table

```
                    3.78986544731362 3.79586172862875

                    3.79861060969518 3.80082840189958

                    3.80611993804014 3.81431756809632

                    3.81645379756594 3.81907245004588

                    3.82012253828179 3.83547896264376

                    3.8395732813918 3.84005689660828

                    3.84389018851706 3.84682632206803

                    3.8509692894115 3.85250128932439 3.871100451953

                    3.88595319563806 3.89383049131859
```

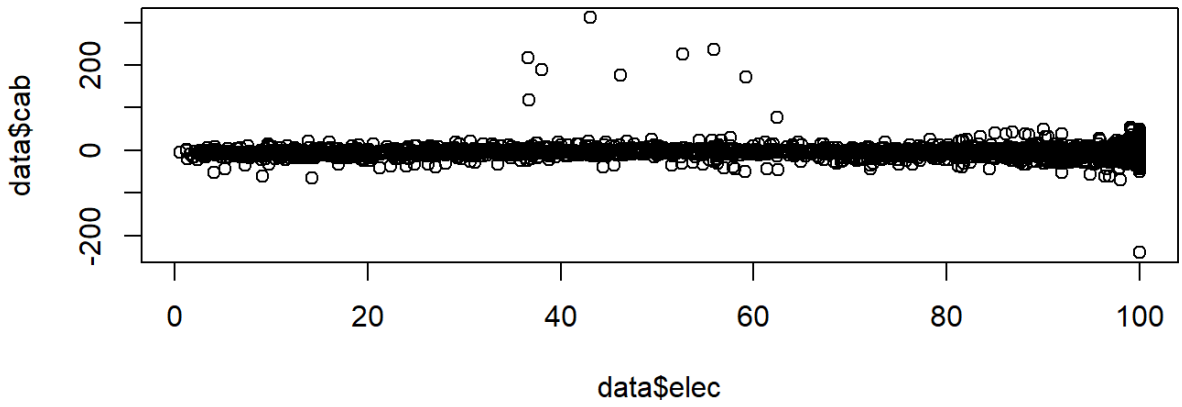


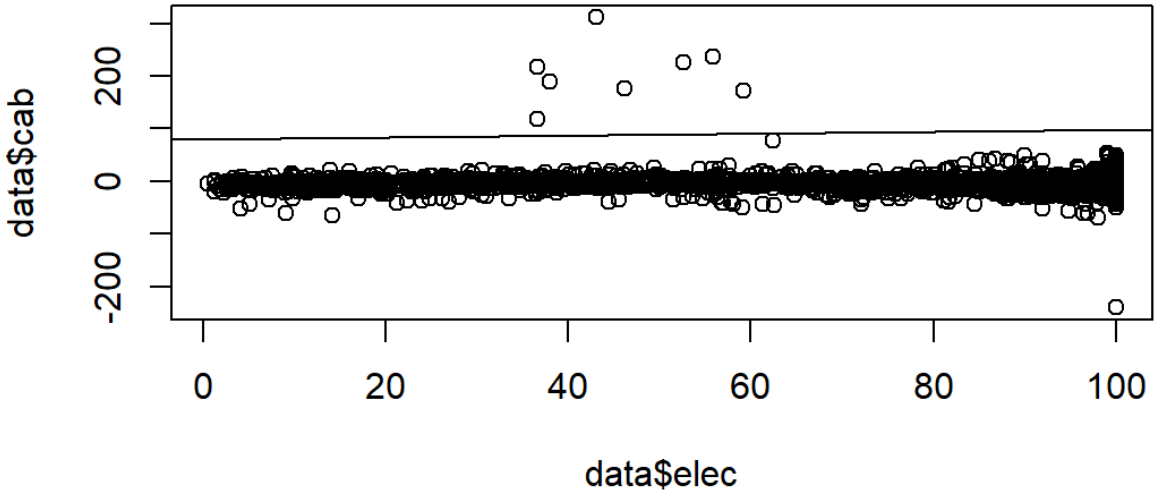


```
# Correlation and Regression
#Variance
v1=var(data$elec,na.rm=TRUE); v1
v2=var(data$cab,na.rm=TRUE); v2
s11=sqrt(v1); s11
s21=sqrt(v2); s21
s1=sd(data$elec,na.rm=TRUE);s1
s2=sd(data$cab,na.rm=TRUE);s2
corr=cor(data$elec,data$cab); corr
# Covariance between "elec" and "cab"
covariance=cov(data$elec,data$cab); covariance
r=corr/(s1&s2);r # Karl pearson's coefficient
cd=corr*corr; cd
# Visualize the samples
plot(data$elec,data$cab)
regression=lm(data$elec~data$cab); regression
abline(regression)
summary(regression)
```

```
> # Correlation and Regression
> #Variance
> v1=var(data$elec,na.rm=TRUE); v1
[1] 835.3094
> v2=var(data$cab,na.rm=TRUE); v2
[1] 172.0246
> s11=sqrt(v1); s11
[1] 28.90172
> s21=sqrt(v2); s21
[1] 13.11582
> s1=sd(data$elec,na.rm=TRUE);s1
[1] 28.90172
> s2=sd(data$cab,na.rm=TRUE);s2
[1] 13.11582
> corr=cor(data$elec,data$cab); corr
[1] NA

> # Covariance between "elec" and "cab"
> covariance=cov(data$elec,data$cab); covariance
[1] NA
> r=corr/(s1&s2);r # Karl pearson's coefficient
[1] NA
> cd=corr*corr; cd
[1] NA
```

```
# Visaulize the samples
```





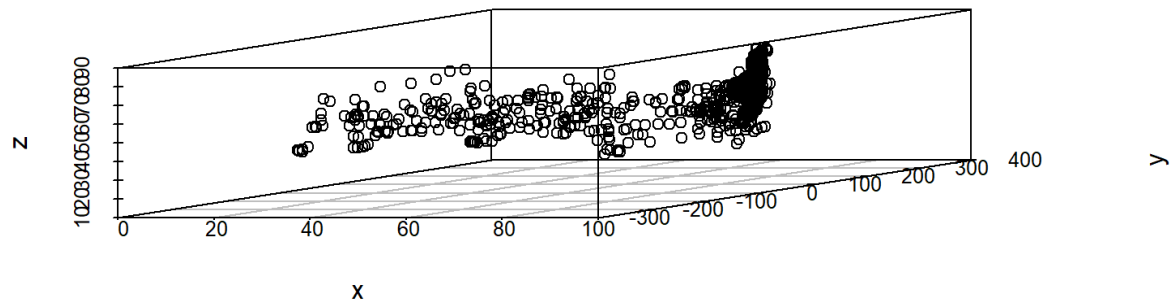
```
> regression=lm(data$elec~data$cab); regression

Call:
lm(formula = data$elec ~ data$cab)

Coefficients:
(Intercept)     data$cab
    80.9018       0.1689

> abline(regression)
> summary(regression)

Call:
lm(formula = data$elec ~ data$cab)

Residuals:
   Min     1Q Median     3Q    Max
-90.52 -11.11  16.90  19.27  59.72

Coefficients:
            Estimate Std. Error t value
(Intercept) 80.90180    0.43810 184.667
data$cab     0.16888    0.03035   5.564
            Pr(>|t|)
(Intercept)  < 2e-16 ***
data$cab    2.78e-08 ***
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
  0.1 ' ' 1

Residual standard error: 29.11 on 4548 degrees of freedom
  (11676 observations deleted due to missingness)
Multiple R-squared:  0.006761,  Adjusted R-squared:  0.006543
F-statistic: 30.96 on 1 and 4548 DF,  p-value: 2.785e-08

>
```

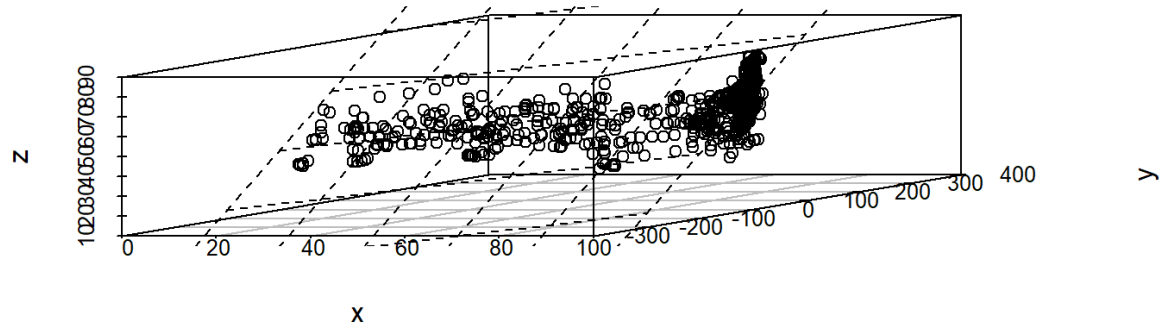
 # Multiple regression

```
x=data$elec; x

y=data$cab; y

z=data$edb; z

reg1=lm(z~x+y); reg1

summary(reg1);

library(scatterplot3d)

graph=scatterplot3d(x,y,z)

graph$plane3d(reg1)
```

```
> # Multiple regression
> x=data$elec; x
   [1]          NA          NA          NA          NA          NA
   [6]          NA          NA          NA          NA          NA
  [11]          NA          NA          NA          NA          NA
  [16]          NA          NA          NA          NA          NA
  [21]          NA          NA          NA          NA          NA
  [26]          NA          NA          NA          NA          NA
  [31]          NA          NA          NA          NA          NA
  [36]          NA          NA          NA          NA          NA
  [41]    1.613591    4.074574    9.409158   14.738506   20.064968
  [46]   25.390894   30.718691   36.051010   42.400002   46.740051
  [51]   42.700001   43.222019   69.099998   68.290649   89.500000
  [56]   71.500000   97.699997   97.699997   96.616135   97.699997
  [61]   97.699997          NA          NA          NA          NA
  [66]          NA          NA          NA          NA          NA
  [71]          NA          NA          NA          NA          NA
  [76]          NA          NA          NA          NA          NA

> y=data$cab; y
   [1]             NA             NA             NA             NA
   [5]             NA             NA             NA             NA
   [9]             NA             NA             NA             NA
  [13]             NA             NA             NA             NA
  [17]             NA             NA             NA  -2.693391157
  [21]   1.471830649  -5.915853607             NA             NA
  [25]             NA             NA             NA             NA
  [29]             NA             NA             NA             NA
  [33]             NA             NA             NA             NA
  [37]             NA             NA             NA             NA
  [41]             NA             NA             NA             NA
  [45]             NA             NA             NA             NA
  [49]  -2.357988549   2.236017922  -3.643311575 -12.619527639
  [53] -25.870698017 -25.290073329 -15.772421201 -21.912668651
  [57] -14.950203387 -18.955947953 -21.585267072 -20.170457679

> z=data$edb; z
   [1]       NA       NA       NA       NA       NA       NA       NA
   [8]       NA       NA       NA       NA       NA       NA       NA
  [15]       NA       NA       NA       NA       NA       NA       NA
  [22]       NA       NA       NA       NA       NA       NA       NA
  [29]       NA       NA       NA       NA       NA       NA       NA
  [36]       NA       NA       NA       NA       NA       NA       NA
  [43]       NA       NA       NA       NA       NA       NA       NA
  [50]       NA       NA       NA       NA       NA       NA 39.25519
  [57] 38.93563 37.13062 44.20343 44.06497       NA       NA       NA
  [64]       NA       NA       NA       NA       NA       NA       NA
  [71]       NA       NA       NA       NA       NA       NA       NA
  [78]       NA       NA       NA       NA       NA       NA       NA
  [85]       NA       NA       NA       NA       NA       NA       NA
  [92]       NA       NA       NA       NA       NA       NA       NA
  [99]       NA       NA       NA       NA       NA       NA       NA
 [106]       NA       NA       NA       NA       NA       NA       NA
 [113]       NA       NA       NA       NA 49.76015 50.56552 51.72151
```

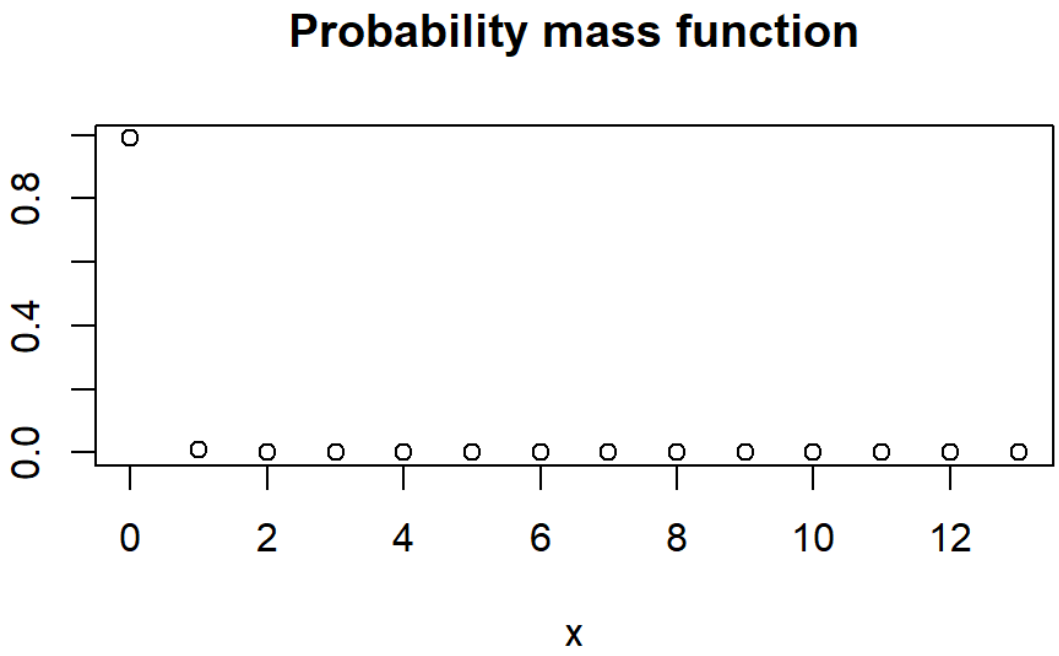
```
> reg1=lm(z~x+y); reg1

Call:
lm(formula = z ~ x + y)

Coefficients:
(Intercept)            x            y
    39.8182       0.2814       0.2550

> 

> summary(reg1);

Call:
lm(formula = z ~ x + y)

Residuals:
     Min       1Q   Median       3Q      Max
-30.6549  -7.0003   0.4658   7.6236  28.3369

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.81821    1.20800  32.962  < 2e-16 ***
x            0.28137    0.01342  20.962  < 2e-16 ***
y            0.25502    0.03842   6.638 5.62e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.894 on 861 degrees of freedom
  (15362 observations deleted due to missingness)
Multiple R-squared:  0.3954,    Adjusted R-squared:  0.394
F-statistic: 281.5 on 2 and 861 DF,  p-value: < 2.2e-16
```

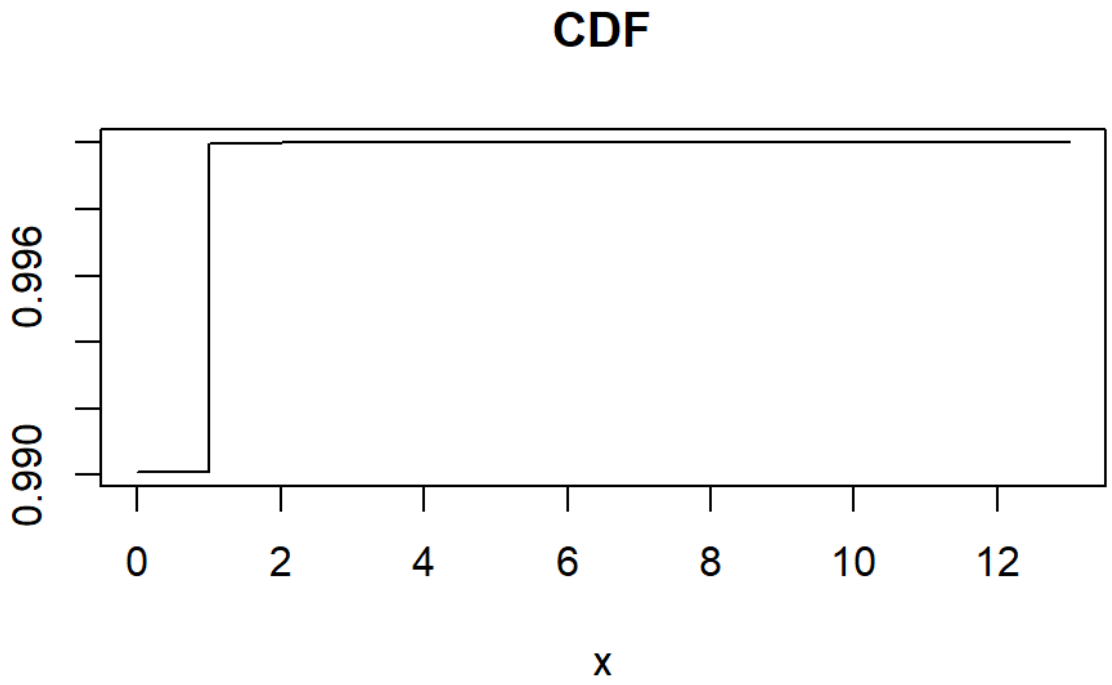





```
# Binomial Distribution

n=13

p=nrow(data[data$elec == "Turkiye" & data$year ==
"2000",])/nrow(data[data$elec == "Turkiye",]);p

dbinom(6,n,p)

#PMF

x = 0:n

pmf = dbinom(x,n,p);pmf

plot(x,pmf,main="Probability mass function");

pbinom(9,n,p);

#CDF

cdf = pbinom(x,n,p);cdf

plot(x,cdf,type = "s",main = "CDF")

mu = n*p;mu

var = n*p*(1-p);var

sd = sqrt(var); sd
```

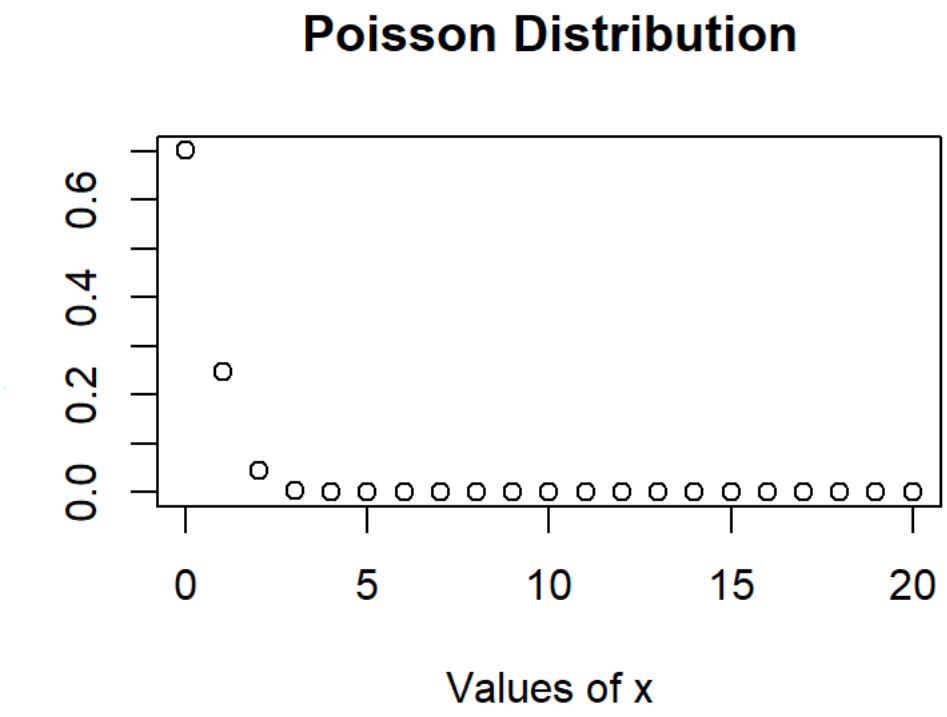
```
> n=13
> p=nrow(data[data$elec == "Turkiye" & data$year == "2000",])/n
row(data[data$elec == "Turkiye",]);p
[1] 0.0007666192
> dbinom(6,n,p)
[1] 3.464687e-16
> #PMF
> x = 0:n
> pmf = dbinom(x,n,p);pmf
 [1] 9.900797e-01 9.874753e-03 4.545590e-05 1.278717e-07
 [5] 2.452603e-10 3.386979e-13 3.464687e-16 2.658133e-19
 [9] 1.529505e-22 6.519151e-26 2.000616e-29 4.186057e-33
[13] 5.352623e-37 3.158901e-41
> plot(x,pmf,main="Probability mass function");
> pbinom(9,n,p);
[1] 1
> #CDF
> cdf = pbinom(x,n,p);cdf
 [1] 0.9900797 0.9999544 0.9999999 1.0000000 1.0000000
 [6] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
[11] 1.0000000 1.0000000 1.0000000 1.0000000
> plot(x,cdf,type = "s",main = "CDF")
> mu = n*p;mu
[1] 0.00996605
> var = n*p*(1-p);var
[1] 0.00995841
> sd = sqrt(var); sd
[1] 0.09979183
>
```






```
# Poisson Distribution

n=20;n

ps=nrow(data[data$edb == "Afghanistan" & data$year ==
"2014",])/nrow(data[data$edb == "Afghanistan",]);ps

lambda=n*ps;lambda

xn=0:n

pxn=dpois(xn,lambda);pxn

plot(xn,pxn,type="p",xlab="Values of x",ylab="Probability
distribution of x",main="Poisson Distribution")

Ex=weighted.mean(xn,pxn);Ex

Var=weighted.mean(xn*xn,pxn)-Ex^2;Var
```

```
> #Poisson Distribution-X
> n=20;n
[1] 20
> ps=nrow(data[data$edb == "Afghanistan" & data$year == "201
4",])/nrow(data[data$edb == "Afghanistan",]);ps
[1] 0.01768735
> lambda=n*ps;lambda
[1] 0.3537469
> xn=0:n
> pxn=dpois(xn,lambda);pxn
 [1] 7.020526e-01 2.483490e-01 4.392634e-02 5.179602e-03
 [5] 4.580671e-04 3.240797e-05 1.910703e-06 9.655791e-08
 [9] 4.269633e-09 1.678188e-10 5.936539e-12 1.909121e-13
[13] 5.627879e-15 1.531419e-16 3.869535e-18 9.125573e-20
[17] 2.017590e-21 4.198330e-23 8.250813e-25 1.536158e-26
[21] 2.717055e-28
> plot(xn,pxn,type="p",xlab="Values of x",ylab="Probability di
stribution of x",main="Poisson Distribution")
> Ex=weighted.mean(xn,pxn);Ex
[1] 0.3537469
> Var=weighted.mean(xn*xn,pxn)-Ex^2;Var
[1] 0.3537469
>
```



 # Hypothesis Testing

```
mu = nrow(data[data$cpi == "Afghanistan" & data$year ==
"2014",])/nrow(data[data$cpi == "Afghanistan",]);mu

n = 35;

x_bar = nrow(data[data$cpi == "Turkiye" & data$year ==
"2000",])/nrow(data[data$cpi == "Turkiye",]);x_bar

sig = 2.5;

alpha = 0.05;

z = (x_bar-mu)/(sig/sqrt(n));z; #test statistic

#two tailed critical value

zhalfalpha = qnorm(1-(alpha/2));zhalfalpha;

#qnorm takes the cumulative probability and gives the corresponding
z-value

c(-zhalfalpha,zhalfalpha); #vector representation

#comparison

if(-(zhalfalpha)<z | z<zhalfalpha){print("Accept Null
hypothesis")}else{print("Reject Null hypothesis")}
```

Giving assumed values for sigma and alpha:

```
> # Hypothesis Testing
> mu = nrow(data[data$cpi == "Afghanistan" & data$year == "2014",])/nrow
(data[data$cpi == "Afghanistan",]);mu
[1] 0.005567741
> n = 35;
> x_bar = nrow(data[data$cpi == "Turkiye" & data$year == "2000",])/nrow
(data[data$cpi == "Turkiye",]);x_bar
[1] 0.009448287
> sig = 2.5;
> alpha = 0.05;
> z = (x_bar-mu)/(sig/sqrt(n));z; #test statistic
[1] 0.009183049
> #two tailed critical value
> zhalfalpha = qnorm(1-(alpha/2));zhalfalpha;
[1] 1.959964
> #qnorm takes the cumulative probability and gives the corresponding z-
value
> c(-zhalfalpha,zhalfalpha); #vector representation
[1] -1.959964  1.959964
> #comparison

> if(-(zhalfalpha)<z | z<zhalfalpha){print("Accept Null hypothesis")}els
e{print("Reject Null hypothesis")}
[1] "Accept Null hypothesis"
>
```

INFERENCE:

Therefore, a dataset is formed from 5 different data points and is tested under various experiments to get different parameters and graphs.