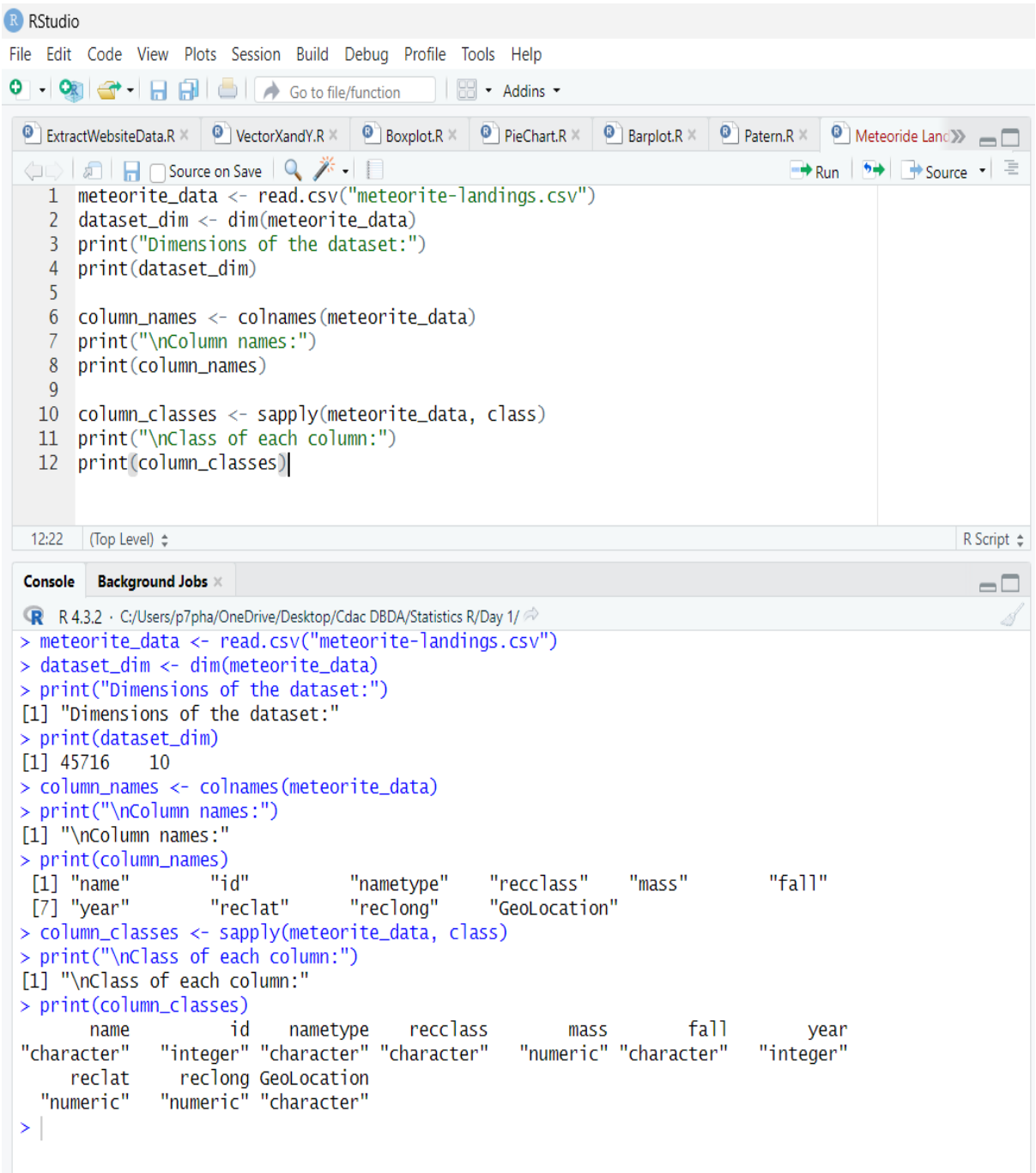
ADVANCED ANALYTICS USING STATISTICS– LAB 1

1. Load the dataset of meterior-landings and do the following using R.

a. Find the dimension, column names, class of each column.

Ans=



b. Select name, recclass, mass, reclat, reclong from the dataframe.

Ans=

```
14 selected_columns <- meteorite_data[, c("name", "recclass", "mass", "reclat", "reclong")]
15 print("\nselected_columns of dataset: ")
16 print(selected_columns)
```

Console Background Jobs

R 4.3.2 · C:/Users/p7pha/OneDrive/Desktop/Cdac DBDA/Statistics R/Day 1/

```
> selected_columns <- meteorite_data[, c("name", "recclass", "mass", "reclat", "reclong")]
> print("\nselected_columns of dataset: ")
[1] "\nselected_columns of dataset: "
> print(selected_columns)
```

	name	recclass	mass	reclat	reclong
1	Aachen	L5	21.0	50.77500	6.08333
2	Aarhus	H6	720.0	56.18333	10.23333
3	Abee	EH4	107000.0	54.21667	-113.00000
4	Acapulco	Acapulcoite	1914.0	16.88333	-99.90000
5	Achiras	L6	780.0	-33.16667	-64.95000
6	Adhi Kot	EH4	4239.0	32.10000	71.80000
7	Adzhi-Bogdo (stone)	LL3-6	910.0	44.83333	95.16667
8	Agen	H5	30000.0	44.21667	0.61667
9	Aguada	L6	1620.0	-31.60000	-65.23333
10	Aguila Blanca	L	1440.0	-30.86667	-64.55000
11	Aioun el Atrouss	Diogenite-pm	1000.0	16.39806	-9.57028
12	Aïr	L6	24000.0	19.08333	8.38333
13	Aire-sur-la-Lys	Unknown	NA	50.66667	2.33333
14	Akaba	L6	779.0	29.51667	35.05000
15	Akbarpur	H4	1800.0	29.71667	77.95000
16	Akwanga	H	3000.0	8.91667	8.43333
17	Akyumak	Iron, IVA	50000.0	39.91667	42.81667
18	Al Rais	CR2-an	160.0	24.41667	39.51667
19	Al Zarnkh	LL5	700.0	13.66033	28.96000

c. Check whether there are null values. If there are null values, give the count of null values.

Ans=

```
17
18 library(tidyverse)
19
20 null_counts <- meteorite_data %>% summarise_all(~sum(is.na(.))) %>% gather() %>%filter(value > 0)
21
22 if (nrow(null_counts) > 0) {
23   print("There are null values in the dataset.")
24   print(null_counts)
25 } else {
26   print("There are no null values in the dataset.")
27 }
```

27:2 (Top Level) R Script

Console Background Jobs

R 4.3.2 · C:/Users/p7pha/OneDrive/Desktop/Cdac DBDA/Statistics R/Day 1/

```
> library(tidyverse)
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✓ dplyr     1.1.3     ✓ readr     2.1.4
✓ forcats   1.0.0     ✓ stringr   1.5.0
✓ ggplot2   3.4.4     ✓ tibble    3.2.1
✓ lubridate 1.9.3     ✓ tidyr     1.3.0
✓ purrr     1.0.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✗ dplyr::filter() masks stats::filter()
✗ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package to force all conflicts to become errors
> null_counts <- meteorite_data %>% summarise_all(~sum(is.na(.))) %>% gather() %>%filter(value > 0)
> if (nrow(null_counts) > 0) {
+   print("There are null values in the dataset.")
+   print(null_counts)
+ } else {
+   print("There are no null values in the dataset.")
+ }
[1] "There are null values in the dataset."
      key value
1    mass   131
2    year   288
3  reclat  7315
4 reclong  7315
>
```

d. Omit the rows with null values and store in a dataframe.

Ans=

```
29  cleaned_data <- na.omit(meteorite_data)
30  print("data after omitting rows with null values:")
31  print(head(cleaned_data))
```

```
> cleaned_data <- na.omit(meteorite_data)
> print("data after omitting rows with null values:")
[1] "data after omitting rows with null values:"
> print(head(cleaned_data))
     name  id nametype    recclass   mass fall year    reclat    reclong            GeoLocation
1   Aachen   1    Valid          L5     21 Fell 1880  50.77500    6.08333    (50.775000, 6.083330)
2   Aarhus   2    Valid          H6    720 Fell 1951  56.18333   10.23333   (56.183330, 10.233330)
3     Abee   6    Valid         EH4 107000 Fell 1952  54.21667 -113.00000 (54.216670, -113.000000)
4 Acapulco  10    Valid Acapulcoite   1914 Fell 1976  16.88333  -99.90000  (16.883330, -99.900000)
5  Achiras 370    Valid          L6    780 Fell 1902 -33.16667  -64.95000 (-33.166670, -64.950000)
6 Adhi Kot 379    Valid         EH4   4239 Fell 1919  32.10000   71.80000   (32.100000, 71.800000)
>
```

e. Display only those rows with recclass = "L5" and mass > 100.

Ans=

```
33 library(dplyr)
34
35 filtered_data <- meteorite_data %>%filter(recclass == "L5", mass > 100)
36 print("Filtered data with recclass = 'L5' and mass > 100:")
37 print(filtered_data)
```

37:21 (Top Level) R Script

Console Background Jobs

R 4.3.2 · C:/Users/p7pha/OneDrive/Desktop/Cdac DBDA/Statistics R/Day 1/

```
> library(dplyr)
> filtered_data <- meteorite_data %>%filter(recclass == "L5", mass > 100)
> print("Filtered data with recclass = 'L5' and mass > 100:")
[1] "Filtered data with recclass = 'L5' and mass > 100:"
> print(filtered_data)
                    name    id nametype recclass      mass  fall year    reclat    reclong
1   Northwest Africa 5815 50693    Valid       L5    256.80 Found   NA   0.00000    0.00000
2                 Ausson  4903    Valid       L5  50000.00  Fell 1858  43.08333    0.58333
3                Barwell  4954    Valid       L5  44000.00  Fell 1965  52.56528   -1.33972
4              Baszkówka  4957    Valid       L5  15500.00  Fell 1994  52.03333   20.93583
5                 Beuste  5034    Valid       L5   2000.00  Fell 1859  43.21667   -0.23333
6    Black Moshannan Park  5065    Valid       L5    705.00  Fell 1941  40.91667  -78.08333
7               Blackwell  5068    Valid       L5   2381.00  Fell 1906  36.83333  -97.33333
8                  Borkut  5113    Valid       L5   7000.00  Fell 1852  48.15000   24.28333
9            Campos Sales  5249    Valid       L5  23680.00  Fell 1991  -7.03333  -40.16667
10                Chajari  5316    Valid       L5  18300.00  Fell 1933 -30.78333  -58.05000
11             Chandakapur  5320    Valid       L5   8800.00  Fell 1838  20.26667   76.01667
12              Chervettaz  5341    Valid       L5    705.00  Fell 1901  46.55000    6.81667
13                 Cilimus  5364    Valid       L5   1600.00  Fell 1979  -6.95000  108.10000
14                 Crumlin  5477    Valid       L5   4255.00  Fell 1902  54.61667   -6.21667
15                   Daule 51559    Valid       L5   6580.00  Fell 2008  -1.87089  -79.95756
16               Domanitch  7661    Valid       L5    438.00  Fell 1907  40.00000   29.00000
17                Elenovka  7824    Valid       L5  54640.00  Fell 1951  47.83333   37.66667
18                  Ergheo 10044    Valid       L5  20000.00  Fell 1889   1.16667   44.16667
19              Farmington 10074    Valid       L5  89400.00  Fell 1890  39.75000  -97.03333
20                    Fuhe 52412    Valid       L5  23000.00  Fell 1945  31.47556  113.56694
21                Fukutomi 10836    Valid       L5  11620.00  Fell 1882  33.18333  130.20000
22                  Guibga 11442    Valid       L5    288.00  Fell 1972  13.50000   -0.68333
23               Homestead 11901    Valid       L5 230000.00  Fell 1875  41.80000  -91.86667
24                Honolulu 11904    Valid       L5   2420.00  Fell 1825  21.30000 -157.86667
25               Innisfree 12039    Valid       L5   4576.00  Fell 1977  53.41500 -111.33750
```

```
70     (27.733330, 4.400000)
71     (27.616670, 4.416670)
72     (27.616670, 4.566670)
73     (27.716670, 4.133330)
74     (27.766670, 4.533330)
75     (27.666670, 4.083330)
76     (27.566670, 4.083330)
77     (27.516670, 4.000000)
78     (27.640280, 4.170560)
79     (27.573890, 4.119720)
80     (27.542220, 3.884440)
81     (27.490830, 3.897780)
82     (27.578610, 3.968890)
83     (27.583330, 4.300000)
84     (27.979330, 4.278170)
85     (27.766670, 4.016670)
86     (27.609330, 3.935670)
87     (27.678170, 4.465330)
88     (27.360170, 3.700000)
89   (32.590330, -101.772170)
90   (-67.183330, 142.383330)
91     (27.816670, 0.133330)
92     (27.569720, 4.525560)
93     (27.578060, 4.322500)
94     (27.591670, 4.316940)
95    (19.544820, 57.098070)
96    (19.841280, 57.009530)
97      (0.000000, 0.000000)
98      (0.000000, 0.000000)
99   (-77.050740, 157.198340)
100 (-76.711580, 158.774530)
 [ reached 'max' / getOption("max.print") -- omitted 1461 rows ]
>
```

f. Arrange the data in the descending order of year and display only name, recclass, mass and year.

Ans=

```
39 arranged_data <- meteorite_data %>% arrange(desc(year)) %>% select(name, recclass, mass, year)
40 print("Arranged data in descending order of year: ")
41 print(arranged_data)
```

41:21 (Top Level) R Script

Console Background Jobs

R 4.3.2 · C:/Users/p7pha/OneDrive/Desktop/Cdac DBDA/Statistics R/Day 1/

```
> arranged_data <- meteorite_data %>% arrange(desc(year)) %>% select(name, recclass, mass, year)
> print("Arranged data in descending order of year: ")
[1] "Arranged data in descending order of year: "
> print(arranged_data)
```

	name	recclass	mass	year
1	Ur	Iron	NA	2501
2	Northwest Africa 7701	CK6	55.00	2101
3	Chelyabinsk	LL5	100000.00	2013
4	Northwest Africa 7755	Martian (shergottite)	30.00	2013
5	Northwest Africa 7812	Angrite	46.20	2013
6	Northwest Africa 7822	Achondrite-ung	45.80	2013
7	Northwest Africa 7855	H4	916.00	2013
8	Northwest Africa 7856	LL6	517.00	2013
9	Northwest Africa 7857	LL6	246.00	2013
10	Northwest Africa 7858	H4	459.00	2013
11	Northwest Africa 7861	L5	611.00	2013
12	Northwest Africa 7862	L4/5	317.00	2013
13	Northwest Africa 7863	LL5	1000.00	2013
14	Battle Mountain	L6	2900.00	2012
15	Sutter's Mill	C	992.50	2012
16	Antelope	H4	754.00	2012
17	Catalina 009	CR2	5.20	2012
18	Jiddat al Harasis 799	LL6	212.00	2012
19	Johannesburg	H4	63.00	2012
20	Ksar Ghilane 010	L5	50.10	2012
21	Ksar Ghilane 011	L4	25.60	2012
22	Los Vientos 028	H~5	12110.00	2012
23	Mandalay Spring	L6	2854.00	2012
24	Northwest Africa 7035	Eucrite-mmict	816.00	2012
25	Northwest Africa 7191	L-melt rock	137.00	2012
26	Northwest Africa 7192	LL4	1780.00	2012
27	Northwest Africa 7194	R4	141.60	2012
28	Northwest Africa 7212	H4	301.30	2012
29	Northwest Africa 7249	L5	6760.00	2012
30	Northwest Africa 7250	LL6	816.00	2012

R 4.3.2 · C:/Users/p7pha/OneDrive/Desktop/Cdac DBDA/Statistics R/Day 1/

```
218 Northwest Africa 7770                      H5   4920.00 2012
219 Northwest Africa 7771                      H5    220.50 2012
220 Northwest Africa 7772                      L5     29.20 2012
221 Northwest Africa 7773                      H4     33.20 2012
222 Northwest Africa 7776                      L5   2073.00 2012
223 Northwest Africa 7777                    H3.8   1352.30 2012
224 Northwest Africa 7778                   LL4-6    235.80 2012
225 Northwest Africa 7779                 Eucrite     49.30 2012
226 Northwest Africa 7780                 Eucrite     26.90 2012
227 Northwest Africa 7781                      L4    646.90 2012
228 Northwest Africa 7782                     LL4    127.80 2012
229 Northwest Africa 7783                      H6    190.20 2012
230 Northwest Africa 7784                      H5    299.70 2012
231 Northwest Africa 7785                      L6    600.70 2012
232 Northwest Africa 7786                     LL6    298.10 2012
233 Northwest Africa 7825                Diogenite     20.15 2012
234 Northwest Africa 7826                     LL6     30.30 2012
235 Northwest Africa 7837                     CR2    586.00 2012
236 Northwest Africa 7839                     LL5   2300.00 2012
237 Northwest Africa 7841                      L6   1070.00 2012
238 Northwest Africa 7842                      H6    365.00 2012
239 Northwest Africa 7843                      H5    236.00 2012
240 Northwest Africa 7850                      H5    625.00 2012
241 Northwest Africa 7851                      H4    585.00 2012
242 Northwest Africa 7852                      H5     39.00 2012
243 Northwest Africa 7860                      H6    500.00 2012
244       Österplana 049               Relict OC      0.00 2012
245             Rosamond                     LL3     11.10 2012
246   Stewart Valley 012                      H6    130.00 2012
247               Tupelo                     EL6    280.00 2012
248      Boumdeid (2011)                      L6   3599.00 2011
249            Sołtmany                       L6   1066.00 2011
250               Thika                       L6  14200.00 2011
 [ reached 'max' / getOption("max.print") -- omitted 45466 rows ]
> 
```

g. Scale the data in the mass variable.

Ans=

h. Find the numerical variables and categorical variables.

Ans=

Numerical variables :-

- id
- mass
- year
- reclat
- reclong

Categorical variable :-

- name
- nametype
- recclass
- fall
- GeoLocation

i. Find the ordinal and nominal data.

Ans=

Ordinal data:-

- Name
- Nametype
- fall

Nominal data:-

- id
- recclass
- mass
- year
- reclat
- reclong
- GeoLocation

j. Find the discrete and continuous data.

Ans=

Discrete data:-

- id
- mass
- year

Continuous data:-

- reclat
- reclong
- GeoLocation

k. Group the data according to Fall variable.

Ans=

```
grouped_data <- meteorite_data %>%group_by(fall)
print(grouped_data)
```

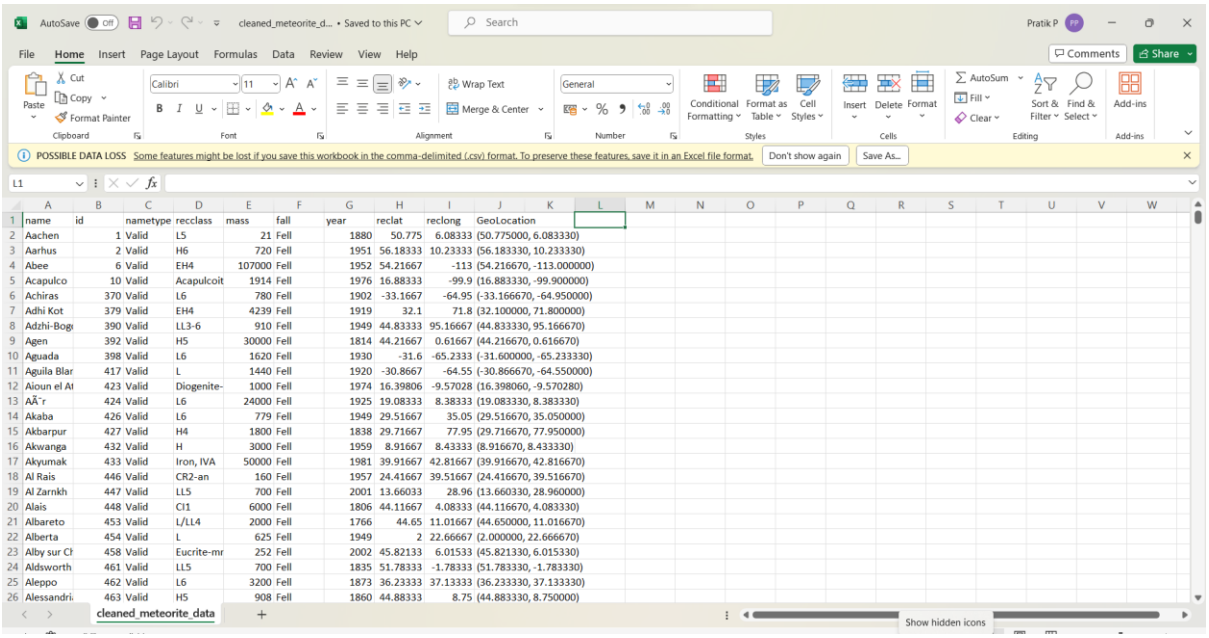
```
> grouped_data <- meteorite_data %>%group_by(fall)
> print(grouped_data)
# A tibble: 45,716 × 10
# Groups:   fall [2]
   name                  id nametype recclass      mass fall   year reclat  reclong GeoLocation
   <chr>              <int> <chr>    <chr>        <dbl> <chr> <int>  <dbl>    <dbl> <chr>
 1 Aachen                 1 Valid    L5              21 Fell  1880   50.8     6.08  (50.775000, 6.0833…
 2 Aarhus                 2 Valid    H6             720 Fell  1951   56.2    10.2   (56.183330, 10.233…
 3 Abee                   6 Valid    EH4         107000 Fell  1952   54.2  -113     (54.216670, -113.0…
 4 Acapulco              10 Valid    Acapulcoite    1914 Fell  1976   16.9   -99.9   (16.883330, -99.90…
 5 Achiras              370 Valid    L6              780 Fell  1902  -33.2   -65.0   (-33.166670, -64.9…
 6 Adhi Kot             379 Valid    EH4            4239 Fell  1919   32.1    71.8   (32.100000, 71.800…
 7 Adzhi-Bogdo (stone)  390 Valid    LL3-6           910 Fell  1949   44.8    95.2   (44.833330, 95.166…
 8 Agen                 392 Valid    H5            30000 Fell  1814   44.2     0.617 (44.216670, 0.6166…
 9 Aguada               398 Valid    L6             1620 Fell  1930  -31.6   -65.2   (-31.600000, -65.2…
10 Aguila Blanca        417 Valid    L              1440 Fell  1920  -30.9   -64.6   (-30.866670, -64.5…
# i 45,706 more rows
```

l. Store the dataframe after cleaning, into a CSV file.

Ans=

```
write.csv(cleaned_data, file = "cleaned_meteorite_data.csv", row.names = FALSE)
```

```
> write.csv(cleaned_data, file = "cleaned_meteorite_data.csv", row.names = FALSE)
>
```



	A	B	C	D	E	F	G	H	I	J	K
1	name	id	nametype	recclass	mass	fall	year	reclat	reclong	GeoLocation	
2	Aachen	1	Valid	L5	21	Fell	1880	50.775	6.08333	(50.775000, 6.083330)	
3	Aarhus	2	Valid	H6	720	Fell	1951	56.18333	10.23333	(56.183330, 10.233330)	
4	Abee	6	Valid	EH4	107000	Fell	1952	54.21667	-113	(54.216670, -113.000000)	
5	Acapulco	10	Valid	Acapulcoit	1914	Fell	1976	16.88333	-99.9	(16.883330, -99.900000)	
6	Achiras	370	Valid	L6	780	Fell	1902	-33.1667	-64.95	(-33.166670, -64.950000)	
7	Adhi Kot	379	Valid	EH4	4239	Fell	1919	32.1	71.8	(32.100000, 71.800000)	
8	Adzhi-Bogd	390	Valid	LL3-6	910	Fell	1949	44.83333	95.16667	(44.833330, 95.166670)	
9	Agen	392	Valid	H5	30000	Fell	1814	44.21667	0.61667	(44.216670, 0.616670)	
10	Aguada	398	Valid	L6	1620	Fell	1930	-31.6	-65.2333	(-31.600000, -65.233330)	
11	Aguila Blar	417	Valid	L	1440	Fell	1920	-30.8667	-64.55	(-30.866670, -64.550000)	
12	Aioun el At	423	Valid	Diogenite-	1000	Fell	1974	16.39806	-9.57028	(16.398060, -9.570280)	
13	AÃ¯r	424	Valid	L6	24000	Fell	1925	19.08333	8.38333	(19.083330, 8.383330)	
14	Akaba	426	Valid	L6	779	Fell	1949	29.51667	35.05	(29.516670, 35.050000)	
15	Akbarpur	427	Valid	H4	1800	Fell	1838	29.71667	77.95	(29.716670, 77.950000)	
16	Akwanga	432	Valid	H	3000	Fell	1959	8.91667	8.43333	(8.916670, 8.433330)	
17	Akyumak	433	Valid	Iron, IVA	50000	Fell	1981	39.91667	42.81667	(39.916670, 42.816670)	
18	Al Rais	446	Valid	CR2-an	160	Fell	1957	24.41667	39.51667	(24.416670, 39.516670)	
19	Al Zarnkh	447	Valid	LL5	700	Fell	2001	13.66033	28.96	(13.660330, 28.960000)	
20	Alais	448	Valid	CI1	6000	Fell	1806	44.11667	4.08333	(44.116670, 4.083330)	
21	Albareto	453	Valid	L/LL4	2000	Fell	1766	44.65	11.01667	(44.650000, 11.016670)	
22	Alberta	454	Valid	L	625	Fell	1949	2	22.66667	(2.000000, 22.666670)	
23	Alby sur Ch	458	Valid	Eucrite-mr	252	Fell	2002	45.82133	6.01533	(45.821330, 6.015330)	
24	Aldsworth	461	Valid	LL5	700	Fell	1835	51.78333	-1.78333	(51.783330, -1.783330)	
25	Aleppo	462	Valid	L6	3200	Fell	1873	36.23333	37.13333	(36.233330, 37.133330)	
26	Alessandri	463	Valid	H5	908	Fell	1860	44.88333	8.75	(44.883330, 8.750000)	

2. What are data normalization methods?

Ans=

- *Min-Max Scaling (Normalization):-*

  - Imagine rescaling values on a scale from 0 to 1.
  - Formula: Take the data and transform it so that the minimum value becomes 0 and the maximum value becomes 1.

- *Z-score Standardization (Standard Scaling):-*

  - Think of making the data look like a bell curve with an average of 0 and a spread of 1.
  - Formula: Subtract the average value (mean) from each data point and divide by the standard deviation.

3. What do you meant by KPI?

Ans=

1. *Measurement Tools***:-** KPIs are like meters or gauges used to measure how well something is doing. They give us numbers or values to understand performance.
2. *Tell Us About Progress:-* They are like scorecards showing how close we are to achieving our goals or targets. If they're high, things are going well; if they're low, we might need to change something.
3. *Focused on Important Goals:-* KPIs are linked to the big things we want to achieve. They help us see if we're moving in the right direction or if we need to adjust our strategies.
4. *Help Make Smart Decisions:-* By tracking KPIs regularly, we can see where we're doing great and where we need to improve. This helps us make better decisions to get better results.
5. *Used in Different Areas:-* KPIs are used in many parts of a business, like sales, customer service, finance, and more. They help each area know if they're doing well or if they need to change what they're doing.
6. *Eg:-* In a company, a low "employee turnover rate" KPI means people are staying in their jobs, which can be good for the company.

4. What do you meant by correlation?

Ans=

Correlation is a statistical measure that tells us about the relationship between two variables. It shows how much and in what direction two variables change together. The correlation coefficient ('r') ranges from -1 to +1.

- A value of +1 indicates a perfect positive relationship, where variables move together.
- A value of -1 indicates a perfect negative relationship, where variables move in opposite directions.
- A value of 0 means no relationship exists between the variables. Remember, correlation doesn't imply causation; just because two things are correlated doesn't mean one causes the other. It helps us understand how strongly and in what direction two variables are related. Various types of correlation coefficients exist, such as Pearson, Spearman, and Kendall, each suited for different kinds of relationships.

5. Create a boxplot of the data (10,20,30,40,50,60,70,120), observe with equation whether 120 is outlier or not.

Ans=

```
data <- c(10, 20, 30, 40, 50, 60, 70, 120)
boxplot(data, main = "Boxplot of Data", col = "skyblue", border = "darkblue", boxwex = 0.5, whisklty = 2, staplelty = 1, outcol = "red")

Q1 <- quantile(data, 0.25)
Q3 <- quantile(data, 0.75)
IQR <- Q3 - Q1

lower_cutoff <- Q1 - 1.5 * IQR
upper_cutoff <- Q3 + 1.5 * IQR

if (120 < lower_cutoff | 120 > upper_cutoff) {
  print("120 is an outlier.")
} else {
  print("120 is not an outlier.")
}
```

```
> data <- c(10, 20, 30, 40, 50, 60, 70, 120)
> boxplot(data, main = "Boxplot of Data", col = "skyblue", border = "darkblue", boxwex = 0.5, whisklty = 2, staplelty = 1, outcol = "red")
> Q1 <- quantile(data, 0.25)
> Q3 <- quantile(data, 0.75)
> IQR <- Q3 - Q1
> lower_cutoff <- Q1 - 1.5 * IQR
> upper_cutoff <- Q3 + 1.5 * IQR
> if (120 < lower_cutoff | 120 > upper_cutoff) {
+   print("120 is an outlier.")
+ } else {
+   print("120 is not an outlier.")
+ }
[1] "120 is an outlier."
>
```