```
 # Install the psych package
install.packages("psych")

# Install the pastecs package
install.packages("pastecs")



data <- read.csv("carcinoma.csv")
summary(data)
str(data)
head(data)
tail(data)
library(psych)
describe(data)
library(pastecs)
stat.desc(data)
```

```
  Patient.Id             Age            Gender      Air.Pollution
 Length:1000        Min.   :14.00   Min.   :1.000   Min.   :1.00  
 Class :character   1st Qu.:27.75   1st Qu.:1.000   1st Qu.:2.00  
 Mode  :character   Median :36.00   Median :1.000   Median :3.00  
                    Mean   :37.17   Mean   :1.402   Mean   :3.84  
                    3rd Qu.:45.00   3rd Qu.:2.000   3rd Qu.:6.00  
                    Max.   :73.00   Max.   :2.000   Max.   :8.00  
  Alcohol.use     Dust.Allergy   OccuPational.Hazards  Genetic.Risk 
 Min.   :1.000   Min.   :1.000   Min.   :1.00         Min.   :1.00  
 1st Qu.:2.000   1st Qu.:4.000   1st Qu.:3.00         1st Qu.:2.00  
 Median :5.000   Median :6.000   Median :5.00         Median :5.00  
 Mean   :4.563   Mean   :5.165   Mean   :4.84         Mean   :4.58  
 3rd Qu.:7.000   3rd Qu.:7.000   3rd Qu.:7.00         3rd Qu.:7.00  
 Max.   :8.000   Max.   :8.000   Max.   :8.00         Max.   :7.00  
 chronic.Lung.Disease Balanced.Diet      Obesity         Smoking     
 Min.   :1.00         Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:3.00         1st Qu.:2.000   1st Qu.:3.000   1st Qu.:2.000  
 Median :4.00         Median :4.000   Median :4.000   Median :3.000  
 Mean   :4.38         Mean   :4.491   Mean   :4.465   Mean   :3.948  
 3rd Qu.:6.00         3rd Qu.:7.000   3rd Qu.:7.000   3rd Qu.:7.000  
 Max.   :7.00         Max.   :7.000   Max.   :7.000   Max.   :8.000  
 Passive.Smoker    Chest.Pain    Coughing.of.Blood    Fatigue     
 Min.   :1.000   Min.   :1.000   Min.   :1.000     Min.   :1.000  
 1st Qu.:2.000   1st Qu.:2.000   1st Qu.:3.000     1st Qu.:2.000  
 Median :4.000   Median :4.000   Median :4.000     Median :3.000  
 Mean   :4.195   Mean   :4.438   Mean   :4.859     Mean   :3.856  
 3rd Qu.:7.000   3rd Qu.:7.000   3rd Qu.:7.000     3rd Qu.:5.000  
 Max.   :8.000   Max.   :9.000   Max.   :9.000     Max.   :9.000  
  Weight.Loss    Shortness.of.Breath    Wheezing     Swallowing.Difficulty
 Min.   :1.000   Min.   :1.00        Min.   :1.000   Min.   :1.000        
 1st Qu.:2.000   1st Qu.:2.00        1st Qu.:2.000   1st Qu.:2.000        
 Median :3.000   Median :4.00        Median :4.000   Median :4.000        
 Mean   :3.855   Mean   :4.24        Mean   :3.777   Mean   :3.746        
 3rd Qu.:6.000   3rd Qu.:6.00        3rd Qu.:5.000   3rd Qu.:5.000        
 Max.   :8.000   Max.   :9.00        Max.   :8.000   Max.   :8.000        
 Clubbing.of.Finger.Nails Frequent.Cold     Dry.Cough        Snoring     
 Min.   :1.000            Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:2.000            1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000  
 Median :4.000            Median :3.000   Median :4.000   Median :3.000  
 Mean   :3.923            Mean   :3.536   Mean   :3.853   Mean   :2.926  
 3rd Qu.:5.000            3rd Qu.:5.000   3rd Qu.:6.000   3rd Qu.:4.000  
 Max.   :9.000            Max.   :7.000   Max.   :7.000   Max.   :7.000  
    Level          
 Length:1000       
 Class :character  
 Mode  :character  
```

```
'data.frame':	1000 obs. of  25 variables:
 $ Patient.Id              : chr  "P1" "P10" "P100" "P1000" ...
 $ Age                     : int  33 17 35 37 46 35 52 28 35 46 ...
 $ Gender                  : int  1 1 1 1 1 1 2 2 2 1 ...
 $ Air.Pollution           : int  2 3 4 7 6 4 2 3 4 2 ...
 $ Alcohol.use             : int  4 1 5 7 8 5 4 1 5 3 ...
 $ Dust.Allergy            : int  5 5 6 7 7 6 5 4 6 4 ...
 $ OccuPational.Hazards    : int  4 3 5 7 7 5 4 3 5 2 ...
 $ Genetic.Risk            : int  3 4 5 6 7 5 3 2 6 4 ...
 $ chronic.Lung.Disease    : int  2 2 4 7 6 4 2 3 5 3 ...
 $ Balanced.Diet           : int  2 2 6 7 7 6 2 4 5 3 ...
 $ Obesity                 : int  4 2 7 7 7 7 4 3 5 3 ...
 $ Smoking                 : int  3 2 2 7 8 2 3 1 6 2 ...
 $ Passive.Smoker          : int  2 4 3 7 7 3 2 4 6 3 ...
 $ Chest.Pain              : int  2 2 4 7 7 4 2 3 6 4 ...
 $ Coughing.of.Blood       : int  4 3 8 8 9 8 4 1 5 4 ...
 $ Fatigue                 : int  3 1 8 4 3 8 3 3 1 1 ...
 $ Weight.Loss             : int  4 3 7 2 2 7 4 2 4 2 ...
 $ Shortness.of.Breath     : int  2 7 9 3 4 9 2 2 3 4 ...
 $ Wheezing                : int  2 8 2 1 1 2 2 4 2 6 ...
 $ Swallowing.Difficulty   : int  3 6 1 4 4 1 3 2 4 5 ...
 $ Clubbing.of.Finger.Nails: int  1 2 4 5 2 4 1 2 6 4 ...
 $ Frequent.Cold           : int  2 1 6 6 4 6 2 3 2 2 ...
 $ Dry.Cough               : int  3 7 7 7 2 7 3 4 4 1 ...
 $ Snoring                 : int  4 2 2 5 3 2 4 3 1 5 ...
 $ Level                   : chr  "Low" "Medium" "High" "High" ...
```

	Patient.Id	Age	Gender	Air.Pollution	Alcohol.use	Dust.Allergy	OccuPational.Hazards	Genetic.Risk	chronic.Lung.Disease	Ba
	<chr>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	
1	P1	33	1	2	4	5	4	3	2	
2	P10	17	1	3	1	5	3	4	2	
3	P100	35	1	4	5	6	5	5	4	
4	P1000	37	1	7	7	7	7	6	7	
5	P101	46	1	6	8	7	7	7	6	
6	P102	35	1	4	5	6	5	5	4	

	Patient.Id	Age	Gender	Air.Pollution	Alcohol.use	Dust.Allergy	OccuPational.Hazards	Genetic.Risk	chronic.Lung.Disease
	<chr>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>
995	P994	33	1	6	7	7	7	7	7
996	P995	44	1	6	7	7	7	7	6
997	P996	37	2	6	8	7	7	7	6
998	P997	25	2	4	5	6	5	5	4
999	P998	18	2	6	8	7	7	7	6
1000	P999	47	1	6	5	6	5	5	4

A psych: 25 × 13

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	
	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	
Patient.Id*	1	1000	500.500	288.8194361	500.5	500.50000	370.6500	1	1000	999	0.00000000	-1.20360060	9
Age	2	1000	37.174	12.0054927	36.0	36.43875	13.3434	14	73	59	0.54944374	0.04714549	0
Gender	3	1000	1.402	0.4905473	1.0	1.37750	0.0000	1	2	1	0.39915418	-1.84251545	0
Air.Pollution	4	1000	3.840	2.0303996	3.0	3.84000	2.9652	1	8	7	0.12507550	-1.38914486	0
Alcohol.use	5	1000	4.563	2.6204767	5.0	4.57875	4.4478	1	8	7	-0.01634084	-1.59685675	0
Dust.Allergy	6	1000	5.165	1.9808328	6.0	5.39375	1.4826	1	8	7	-0.64277636	-0.86721005	0
OccuPational.Hazards	7	1000	4.840	2.1078052	5.0	4.95000	2.9652	1	8	7	-0.23380186	-1.36685824	0
Genetic.Risk	8	1000	4.580	2.1269989	5.0	4.65000	2.9652	1	7	6	-0.12628498	-1.59760614	0
chronic.Lung.Disease	9	1000	4.380	1.8485175	4.0	4.41250	2.9652	1	7	6	-0.21980440	-1.30713188	0
Balanced.Diet	10	1000	4.491	2.1355279	4.0	4.53875	2.9652	1	7	6	-0.06430203	-1.64165892	0
Obesity	11	1000	4.465	2.1249212	4.0	4.54375	2.9652	1	7	6	0.02875818	-1.47816326	0
Smoking	12	1000	3.948	2.4959017	3.0	3.82375	2.9652	1	8	7	0.38016896	-1.45299748	0
Passive.Smoker	13	1000	4.195	2.3117784	4.0	4.06875	2.9652	1	8	7	0.41022527	-1.33076732	0
Chest.Pain	14	1000	4.438	2.2802095	4.0	4.43750	2.9652	1	9	8	0.16421338	-1.36189023	0
Coughing.of.Blood	15	1000	4.859	2.4279650	4.0	4.83750	2.9652	1	9	8	0.12163168	-1.29634660	0
Fatigue	16	1000	3.856	2.2446163	3.0	3.64625	1.4826	1	9	8	0.85306470	-0.21892273	0
Weight.Loss	17	1000	3.855	2.2065457	3.0	3.80625	2.9652	1	8	7	0.35406890	-1.39290498	0
Shortness.of.Breath	18	1000	4.240	2.2850868	4.0	4.10000	2.9652	1	9	8	0.40516471	-0.86091362	0
Wheezing	19	1000	3.777	2.0419208	4.0	3.70875	2.9652	1	8	7	0.22348253	-1.18807730	0
Swallowing.Difficulty	20	1000	3.746	2.2703829	4.0	3.55750	2.9652	1	8	7	0.44982417	-0.89324860	0
Clubbing.of.Finger.Nails	21	1000	3.923	2.3880481	4.0	3.67875	2.9652	1	9	8	0.79417567	-0.34627482	0
Frequent.Cold	22	1000	3.536	1.8325016	3.0	3.45875	1.4826	1	7	6	0.40523001	-0.94810789	0
Dry.Cough	23	1000	3.853	2.0390068	4.0	3.81625	2.9652	1	7	6	0.22316407	-1.29377712	0
Snoring	24	1000	2.926	1.4746860	3.0	2.83375	1.4826	1	7	6	0.54839627	-0.55923050	0
Level*	25	1000	1.967	0.8346302	2.0	1.95875	1.4826	1	3	2	0.06180000	-1.56326188	0

	Patient.Id	Age	Gender	Air.Pollution	Alcohol.use	Dust.Allergy	OccuPational.Hazards	Genetic.Risk	chro
	<lgl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	
nbr.val	NA	1.000000e+03	1.000000e+03	1.000000e+03	1.000000e+03	1.000000e+03	1.000000e+03	1.000000e+03	
nbr.null	NA	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	
nbr.na	NA	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	
min	NA	1.400000e+01	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	
max	NA	7.300000e+01	2.000000e+00	8.000000e+00	8.000000e+00	8.000000e+00	8.000000e+00	7.000000e+00	
range	NA	5.900000e+01	1.000000e+00	7.000000e+00	7.000000e+00	7.000000e+00	7.000000e+00	6.000000e+00	
sum	NA	3.717400e+04	1.402000e+03	3.840000e+03	4.563000e+03	5.165000e+03	4.840000e+03	4.580000e+03	
median	NA	3.600000e+01	1.000000e+00	3.000000e+00	5.000000e+00	6.000000e+00	5.000000e+00	5.000000e+00	
mean	NA	3.717400e+01	1.402000e+00	3.840000e+00	4.563000e+00	5.165000e+00	4.840000e+00	4.580000e+00	
SE.mean	NA	3.796470e-01	1.551247e-02	6.420687e-02	8.286675e-02	6.263943e-02	6.665465e-02	6.726161e-02	
CI.mean	NA	7.449971e-01	3.044076e-02	1.259958e-01	1.626129e-01	1.229200e-01	1.307992e-01	1.319902e-01	

var	NA	1.441319e+02	2.406366e-01	4.122523e+00	6.866898e+00	3.923699e+00	4.442843e+00	4.524124e+00
std.dev	NA	1.200549e+01	4.905473e-01	2.030400e+00	2.620477e+00	1.980833e+00	2.107805e+00	2.126999e+00
coef.var	NA	3.229540e-01	3.498911e-01	5.287499e-01	5.742881e-01	3.835107e-01	4.354969e-01	4.644102e-01

```
# Install the dplyr package
install.packages("dplyr")

# Install the tibble package
install.packages("tibble")
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)
```

```
# Load required libraries
library(dplyr) # For data manipulation
library(tibble) # For data manipulation

# Read the CSV file
data <- read.csv("carcinoma.csv")

# Check the dimensions (number of rows and columns) of the dataset
cat("Dimensions of the dataset:\n")
print(dim(data))

# Check the column names and their data types
cat("\nColumn names and data types:\n")
print(sapply(data, class))

# Check the number of missing values in each column
cat("\nNumber of missing values in each column:\n")
print(colSums(is.na(data)))
```

```
Dimensions of the dataset:
[1] 1000   25

Column names and data types:
              Patient.Id                      Age                   Gender 
             "character"                "integer"                "integer" 
           Air.Pollution              Alcohol.use             Dust.Allergy 
               "integer"                "integer"                "integer" 
    OccuPational.Hazards             Genetic.Risk     chronic.Lung.Disease 
               "integer"                "integer"                "integer" 
           Balanced.Diet                  Obesity                  Smoking 
               "integer"                "integer"                "integer" 
          Passive.Smoker               Chest.Pain        Coughing.of.Blood 
               "integer"                "integer"                "integer" 
                 Fatigue              Weight.Loss      Shortness.of.Breath 
               "integer"                "integer"                "integer" 
                Wheezing    Swallowing.Difficulty Clubbing.of.Finger.Nails 
               "integer"                "integer"                "integer" 
           Frequent.Cold                Dry.Cough                  Snoring 
               "integer"                "integer"                "integer" 
                   Level 
             "character" 

Number of missing values in each column:
              Patient.Id                      Age                   Gender 
                       0                        0                        0 
           Air.Pollution              Alcohol.use             Dust.Allergy 
                       0                        0                        0 
    OccuPational.Hazards             Genetic.Risk     chronic.Lung.Disease 
                       0                        0                        0 
           Balanced.Diet                  Obesity                  Smoking 
                       0                        0                        0 
          Passive.Smoker               Chest.Pain        Coughing.of.Blood 
                       0                        0                        0 
                 Fatigue              Weight.Loss      Shortness.of.Breath 
                       0                        0                        0 
                Wheezing    Swallowing.Difficulty Clubbing.of.Finger.Nails 
                       0                        0                        0 
           Frequent.Cold                Dry.Cough                  Snoring 
                       0                        0                        0 
```

```
                         Level
                             0
```

```
install.packages("ggplot2")
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)
```

```
# Check the structure of your dataset
str(data)
```

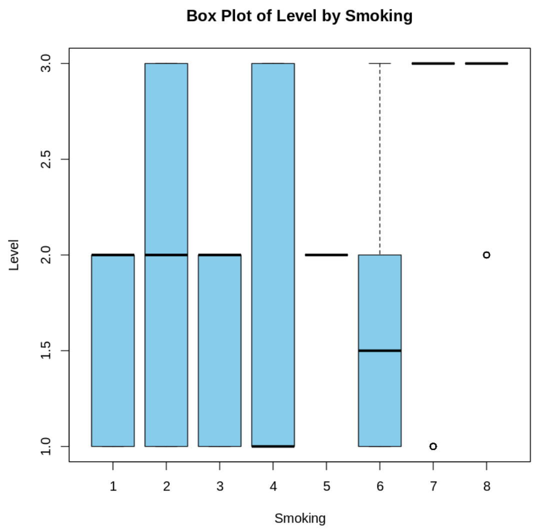
```
'data.frame':   1000 obs. of  25 variables:
 $ Patient.Id              : chr  "P1" "P10" "P100" "P1000" ...
 $ Age                     : int  33 17 35 37 46 35 52 28 35 46 ...
 $ Gender                  : int  1 1 1 1 1 1 2 2 2 1 ...
 $ Air.Pollution           : int  2 3 4 7 6 4 2 3 4 2 ...
 $ Alcohol.use             : int  4 1 5 7 8 5 4 1 5 3 ...
 $ Dust.Allergy            : int  5 5 6 7 7 6 5 4 6 4 ...
 $ OccuPational.Hazards    : int  4 3 5 7 7 5 4 3 5 2 ...
 $ Genetic.Risk            : int  3 4 5 6 7 5 3 2 6 4 ...
 $ chronic.Lung.Disease    : int  2 2 4 7 6 4 2 3 5 3 ...
 $ Balanced.Diet           : int  2 2 6 7 7 6 2 4 5 3 ...
 $ Obesity                 : int  4 2 7 7 7 7 4 3 5 3 ...
 $ Smoking                 : int  3 2 2 7 8 2 3 1 6 2 ...
 $ Passive.Smoker          : int  2 4 3 7 7 3 2 4 6 3 ...
 $ Chest.Pain              : int  2 2 4 7 7 4 2 3 6 4 ...
 $ Coughing.of.Blood       : int  4 3 8 8 9 8 4 1 5 4 ...
 $ Fatigue                 : int  3 1 8 4 3 8 3 3 1 1 ...
 $ Weight.Loss             : int  4 3 7 2 2 7 4 2 4 2 ...
 $ Shortness.of.Breath     : int  2 7 9 3 4 9 2 2 3 4 ...
 $ Wheezing                : int  2 8 2 1 1 2 2 4 2 6 ...
 $ Swallowing.Difficulty   : int  3 6 1 4 4 1 3 2 4 5 ...
 $ Clubbing.of.Finger.Nails: int  1 2 4 5 2 4 1 2 6 4 ...
 $ Frequent.Cold           : int  2 1 6 6 4 6 2 3 2 2 ...
 $ Dry.Cough               : int  3 7 7 7 2 7 3 4 4 1 ...
 $ Snoring                 : int  4 2 2 5 3 2 4 3 1 5 ...
 $ Level                   : chr  "Low" "Medium" "High" "High" ...
```

```
# Convert 'Level' to a factor
data$Level <- factor(data$Level, levels = c("Low", "Medium", "High"))

# Assign numeric values to levels
data$Level <- as.numeric(data$Level)
```

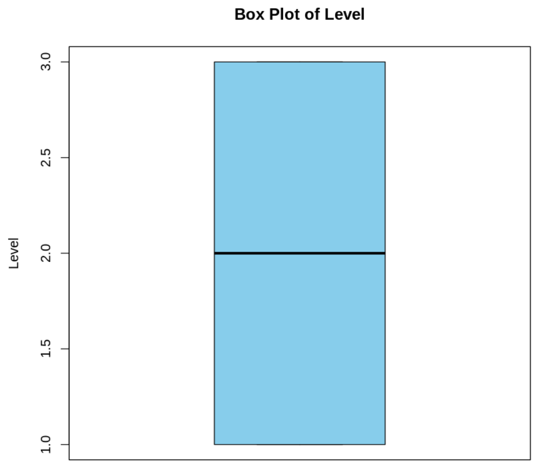
Interpretation:

1.The box plot visually compares the distribution of "Level_numeric" across different smoking categories, providing insights into how smoking status may affect the level of carcinoma. **2.**Comparisons can be made between the median and spread of "Level_numeric" among smokers and non-smokers, aiding in understanding potential associations between smoking behavior and carcinoma level.

```
# Read the dataset
data <- read.csv("carcinoma.csv")

# Convert 'Level' to a factor
data$Level <- factor(data$Level, levels = c("Low", "Medium", "High"))

# Assign numeric values to levels
data$Level_numeric <- as.numeric(data$Level)

# Create the boxplot
boxplot(Level_numeric ~ Smoking, data = data,
        col = "skyblue",  # Set color of the boxes
        main = "Box Plot of Level by Smoking",  # Main title of the plot
        xlab = "Smoking",  # Label for x-axis
        ylab = "Level"     # Label for y-axis
)
```



```
# Read the dataset
data <- read.csv("carcinoma.csv")

# Convert 'Level' to a factor
data$Level <- factor(data$Level, levels = c("Low", "Medium", "High"))

# Create the boxplot for 'Level'
boxplot(data$Level,
        col = "skyblue",  # Set color of the boxes
        main = "Box Plot of Level",  # Main title of the plot
        xlab = "",  # No label for x-axis
        ylab = "Level"  # Label for y-axis
)
```



```
install.packages("corr")
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

Warning message:
“package ‘corr’ is not available for this version of R

A version of this package for your version of R might be available elsewhere,
see the ideas at
https://cran.r-project.org/doc/manuals/r-patched/R-admin.html#Installing-packages”
```

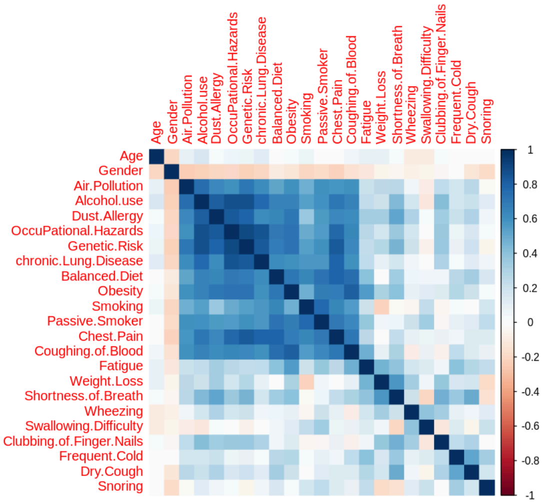
```
# Load dataset
data <- read.csv("carcinoma.csv")
```

```
# Exclude non-numeric variables from the dataset
numeric_data <- subset(carcinoma_data, select = -c(Patient.Id, Level))

# Compute the correlation matrix
correlation_matrix <- cor(numeric_data)
```

```
Error in eval(expr, envir, enclos): object 'carcinoma_data' not found
Traceback:

1. subset(carcinoma_data, select = -c(Patient.Id, Level))
```

```
# Install and load the corrplot package
install.packages("corrplot")
library(corrplot)

# Visualize the correlation matrix
corrplot(correlation_matrix, method = "color")
correlation_matrix
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

corrplot 0.92 loaded
```



```
data <- read.csv("carcinoma.csv")
numeric_data <- data[sapply(data, is.numeric)]
# Step 2: Compute Correlation Matrix
# Calculate the correlation matrix
correlation_matrix <- cor(numeric_data)
correlation_matrix
```

	Age	Gender	Air.Pollution	Alcohol.use	Dust.Allergy	OccuPational.Hazards	Genetic.Risk	chr
Age	1.000000000	-0.2020861307	0.09949419	0.1517417	0.03520170	0.062177375	0.07315054	
Gender	-0.202086131	1.0000000000	-0.24691184	-0.2276359	-0.20431216	-0.192343411	-0.22272747	
Air.Pollution	0.099494194	-0.2469118436	1.00000000	0.7472926	0.63750350	0.608924458	0.70527606	
Alcohol.use	0.151741723	-0.2276359165	0.74729261	1.0000000	0.81864352	0.878785921	0.87720989	
Dust.Allergy	0.035201697	-0.2043121641	0.63750350	0.8186435	1.00000000	0.835859771	0.78790388	
OccuPational.Hazards	0.062177375	-0.1923434108	0.60892446	0.8787859	0.83585977	1.000000000	0.89304852	
Genetic.Risk	0.073150538	-0.2227274663	0.70527606	0.8772099	0.78790388	0.893048523	1.00000000	
chronic.Lung.Disease	0.128951642	-0.2050606213	0.62670091	0.7635758	0.61955592	0.858283853	0.83623083	
Balanced.Diet	0.004863499	-0.0997410643	0.52487291	0.6533519	0.64719683	0.691508821	0.67990485	
Obesity	0.034337163	-0.1238125823	0.60146750	0.6693116	0.70067582	0.722190745	0.72982608	
Smoking	0.075332578	-0.2069242711	0.48190161	0.5470346	0.35869058	0.497692577	0.54325927	
Passive.Smoker	0.004907618	-0.1848261320	0.60676370	0.5925764	0.56000248	0.555310666	0.60907129	
Chest.Pain	0.012863549	-0.2184258147	0.58573351	0.7172423	0.63998312	0.775618729	0.83175083	
Coughing.of.Blood	0.053006399	-0.1465053387	0.60782860	0.6676118	0.62529147	0.645946503	0.63223641	
Fatigue	0.095058772	-0.1164665840	0.21172390	0.2372451	0.33247156	0.267843992	0.23053044	
Weight.Loss	0.106945701	-0.0579932590	0.25801612	0.2078511	0.32175619	0.176225579	0.27174268	
Shortness.of.Breath	0.035329285	-0.0459715849	0.26955773	0.4357853	0.51868168	0.366481599	0.45820047	
Wheezing	-0.095354094	-0.0763038662	0.05536764	0.1808170	0.30485003	0.178925472	0.20497278	
Swallowing.Difficulty	-0.105832694	-0.0583237829	-0.08091767	-0.1140732	0.03114127	-0.002853115	-0.06294835	
Clubbing.of.Finger.Nails	0.039258302	-0.0342191887	0.24106478	0.4149921	0.34571423	0.366446760	0.35781514	
Frequent.Cold	-0.012706476	-0.0005255951	0.17453909	0.1807778	0.21938921	0.077166008	0.08709161	
Dry.Cough	0.012127598	-0.1230008344	0.26148864	0.2112772	0.30019510	0.159887039	0.19439933	
Snoring	-0.004699822	-0.1816184730	-0.02134255	0.1226940	0.05284449	0.022916085	-0.05683068	

```
# Load the dataset from the CSV file
data <- read.csv("carcinoma.csv")

# Exclude non-numeric variables from the dataset
numeric_data <- data[, sapply(data, is.numeric)]

# Compute the correlation matrix
correlation_matrix <- cor(numeric_data)

# Set correlation threshold
cor_threshold <- 0.7

# Find highly correlated features
highly_correlated <- findCorrelation(correlation_matrix, cutoff = cor_threshold)

# Remove redundant features
selected_features <- colnames(correlation_matrix)[-highly_correlated]

# Consider the target variable
# Assuming "Level" is the name of your target variable
# Check correlations with the target variable
cor_with_target <- correlation_matrix["Level", ]
# Select features highly correlated with the target (you can set a threshold here too)
selected_features <- c(selected_features, names(cor_with_target[abs(cor_with_target) > threshold]))

# Validate feature selection
# Perform cross-validation or evaluate model performance metrics
# You can use various modeling techniques or packages for this purpose
# For example, using caret package for cross-validation
library(caret)
# Define your model and cross-validation method
ctrl <- trainControl(method = "cv", number = 5)
# Train your model using the selected features
model <- train(Level ~ ., data = data[, c("Level", selected_features)], method = "lm", trControl = ctrl)
# Evaluate model performance
performance <- summary(model)

# Iterate if necessary
# Iterate through steps 1-5 as needed, adjusting thresholds or criteria based on results
# You may also consider adding additional domain-specific knowledge or techniques for further refinement
```

```
Error in findCorrelation(correlation_matrix, cutoff = cor_threshold): could not find function "findCorrelation"
Traceback:
```

```
install.packages("caret")
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

also installing the dependencies ‘listenv’, ‘parallelly’, ‘future’, ‘globals’, ‘shape’, ‘future.apply’, ‘numDeriv’, ‘progressr’, ‘S
```

```
library(readr)
library(dplyr)
data <- read.csv('carcinoma.csv')
numeric_data <- data[, sapply(data, is.numeric)]
anova_results <- sapply(numeric_data, function(x) {
  if(length(unique(data$Level)) == 2) {
    t_test_result <- t.test(x ~ data$Level)
    p_value <- t_test_result$p.value
  } else {
    anova_result <- aov(x ~ data$Level)
    p_value <- summary(anova_result)[[1]][["Pr(>F)"]][[1]]
  }
  return(p_value)
})
selected_features <- names(sort(anova_results, decreasing = FALSE)[1:10])
print(selected_features)
```

```
Attaching package: ‘dplyr’
```

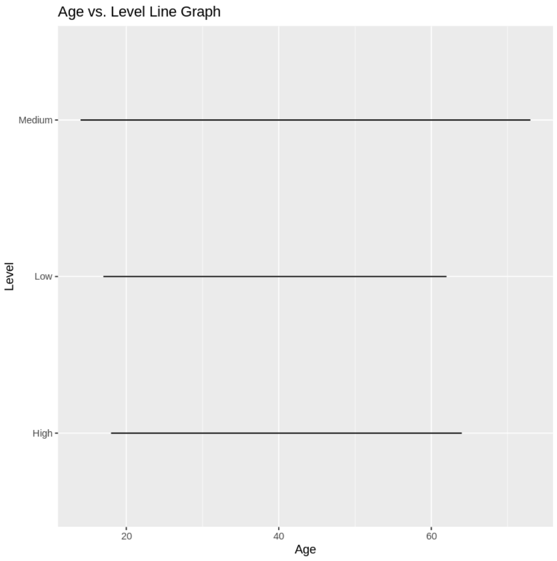
```
The following objects are masked from ‘package:stats’:

    filter, lag


The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union


 [1] "Obesity"              "Coughing.of.Blood"    "Passive.Smoker"      
 [4] "Balanced.Diet"        "Dust.Allergy"         "Alcohol.use"         
 [7] "Genetic.Risk"         "Air.Pollution"        "OccuPational.Hazards"
[10] "Chest.Pain"          
```

```
# Install and load DescTools package
install.packages("DescTools")
library(DescTools)

# Read the dataset
data <- read.csv("carcinoma.csv")

# Display summary statistics
cat("Summary Statistics:\n")
summary(data)

# Calculate mean
cat("\nMean:\n")
means <- sapply(data[, -c(1, 2, 3, 25)], mean)
print(round(means, 2))

# Calculate median
cat("\nMedian:\n")
medians <- sapply(data[, -c(1, 2, 3, 25)], median)
print(medians)

# Calculate mode
cat("\nMode:\n")
modes <- apply(data[, -c(1, 2, 3, 25)], 2, Mode)
print(modes)
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

Summary Statistics:
  Patient.Id             Age            Gender      Air.Pollution
 Length:1000        Min.   :14.00   Min.   :1.000   Min.   :1.00  
 Class :character   1st Qu.:27.75   1st Qu.:1.000   1st Qu.:2.00  
 Mode  :character   Median :36.00   Median :1.000   Median :3.00  
                    Mean   :37.17   Mean   :1.402   Mean   :3.84  
                    3rd Qu.:45.00   3rd Qu.:2.000   3rd Qu.:6.00  
                    Max.   :73.00   Max.   :2.000   Max.   :8.00  
  Alcohol.use     Dust.Allergy   OccuPational.Hazards  Genetic.Risk
 Min.   :1.000   Min.   :1.000   Min.   :1.00         Min.   :1.00  
 1st Qu.:2.000   1st Qu.:4.000   1st Qu.:3.00         1st Qu.:2.00  
 Median :5.000   Median :6.000   Median :5.00         Median :5.00  
 Mean   :4.563   Mean   :5.165   Mean   :4.84         Mean   :4.58  
 3rd Qu.:7.000   3rd Qu.:7.000   3rd Qu.:7.00         3rd Qu.:7.00  
 Max.   :8.000   Max.   :8.000   Max.   :8.00         Max.   :7.00  
 chronic.Lung.Disease Balanced.Diet      Obesity         Smoking     
 Min.   :1.00         Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:3.00         1st Qu.:2.000   1st Qu.:3.000   1st Qu.:2.000  
 Median :4.00         Median :4.000   Median :4.000   Median :3.000  
 Mean   :4.38         Mean   :4.491   Mean   :4.465   Mean   :3.948  
 3rd Qu.:6.00         3rd Qu.:7.000   3rd Qu.:7.000   3rd Qu.:7.000  
 Max.   :7.00         Max.   :7.000   Max.   :7.000   Max.   :8.000  
 Passive.Smoker    Chest.Pain    Coughing.of.Blood    Fatigue     
 Min.   :1.000   Min.   :1.000   Min.   :1.000     Min.   :1.000  
 1st Qu.:2.000   1st Qu.:2.000   1st Qu.:3.000     1st Qu.:2.000  
 Median :4.000   Median :4.000   Median :4.000     Median :3.000  
 Mean   :4.195   Mean   :4.438   Mean   :4.859     Mean   :3.856  
 3rd Qu.:7.000   3rd Qu.:7.000   3rd Qu.:7.000     3rd Qu.:5.000  
 Max.   :8.000   Max.   :9.000   Max.   :9.000     Max.   :9.000  
  Weight.Loss    Shortness.of.Breath    Wheezing     Swallowing.Difficulty
 Min.   :1.000   Min.   :1.00        Min.   :1.000   Min.   :1.000        
 1st Qu.:2.000   1st Qu.:2.00        1st Qu.:2.000   1st Qu.:2.000        
 Median :3.000   Median :4.00        Median :4.000   Median :4.000        
 Mean   :3.855   Mean   :4.24        Mean   :3.777   Mean   :3.746        
 3rd Qu.:6.000   3rd Qu.:6.00        3rd Qu.:5.000   3rd Qu.:5.000        
 Max.   :8.000   Max.   :9.00        Max.   :8.000   Max.   :8.000        
 Clubbing.of.Finger.Nails Frequent.Cold     Dry.Cough        Snoring     
 Min.   :1.000            Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:2.000            1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000  
 Median :4.000            Median :3.000   Median :4.000   Median :3.000  
 Mean   :3.923            Mean   :3.536   Mean   :3.853   Mean   :2.926  
 3rd Qu.:5.000            3rd Qu.:5.000   3rd Qu.:6.000   3rd Qu.:4.000  
 Max.   :9.000            Max.   :7.000   Max.   :7.000   Max.   :7.000  
    Level          
 Length:1000       
 Class :character  
 Mode  :character  
                   
                   
                   

Mean:
           Air.Pollution              Alcohol.use             Dust.Allergy 
                    3.84                     4.56                     5.16 
    OccuPational.Hazards             Genetic.Risk     chronic.Lung.Disease 
                    4.84                     4.58                     4.38 
           Balanced.Diet                  Obesity                  Smoking 
                    4.49                     4.46                     3.95 
          Passive.Smoker               Chest.Pain        Coughing.of.Blood 
                    4.20                     4.44                     4.86 
                 Fatigue              Weight.Loss      Shortness.of.Breath 
                    3.86                     3.86                     4.24 
                Wheezing    Swallowing.Difficulty Clubbing.of.Finger.Nails 
                    3.78                     3.75                     3.92 
           Frequent.Cold                Dry.Cough                  Snoring 
                    3.54                     3.85                     2.93 

Median:
           Air.Pollution              Alcohol.use             Dust.Allergy 
                       3                        5                        6 
    OccuPational.Hazards             Genetic.Risk     chronic.Lung.Disease 
                       5                        5                        4 
           Balanced.Diet                  Obesity                  Smoking 
                       4                        4                        3 
          Passive.Smoker               Chest.Pain        Coughing.of.Blood 
                       4                        4                        4 
                 Fatigue              Weight.Loss      Shortness.of.Breath 
                       3                        3                        4 
                Wheezing    Swallowing.Difficulty Clubbing.of.Finger.Nails 
                       4                        4                        4 
           Frequent.Cold                Dry.Cough                  Snoring 
                       3                        4                        3 

Mode:
           Air.Pollution              Alcohol.use             Dust.Allergy 
                       6                        2                        7 
    OccuPational.Hazards             Genetic.Risk     chronic.Lung.Disease 
```

```
                       7                         7                         6
            Balanced.Diet                   Obesity                   Smoking
                       7                         7                         2
           Passive.Smoker                Chest.Pain         Coughing.of.Blood
                       2                         7                         7
                  Fatigue               Weight.Loss       Shortness.of.Breath
                       3                         2                         2
                 Wheezing    Swallowing.Difficulty Clubbing.of.Finger.Nails
                       2                         1                         2
            Frequent.Cold                 Dry.Cough                   Snoring
                       3                         2                         2
```

```
# Load necessary library
library(ggplot2)

# Assuming your dataset is loaded or created with the name 'data'

# Convert columns to appropriate data types if necessary
data$Age <- as.numeric(data$Age)  # Assuming Age is stored as character initially
data$Level <- as.factor(data$Level)  # Assuming Level is stored as character initially

# Create a line plot for Age against Level
ggplot(data = data, aes(x = Age, y = Level)) +
  geom_line() +  # Adding lines to show trends
  labs(x = "Age", y = "Level") +  # Labeling the axes
  ggtitle("Age vs. Level Line Graph")  # Adding a title
```



Age vs. Level Relationship: The code visualizes the relationship between age and a categorical variable named "Level."

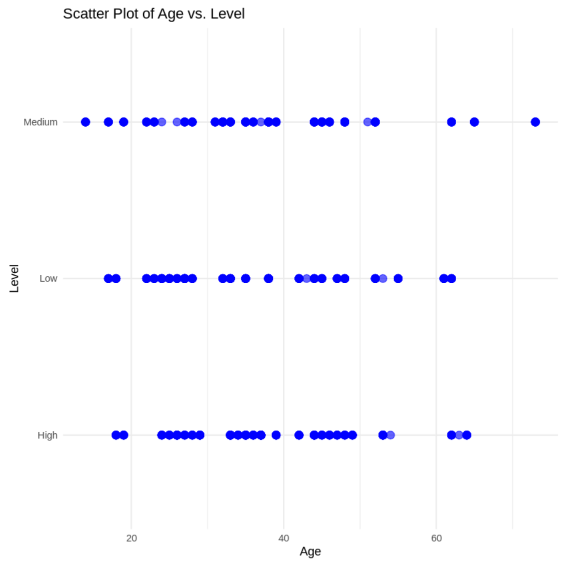
Numeric vs. Categorical Mapping: Age, assumed to be numeric, is mapped to the x-axis, while the categorical variable "Level" is mapped to the y-axis.

Trend Identification: Through the line graph, trends or patterns in the distribution of "Level" across different ages can be discerned.

Age-Related Changes: By observing the line's direction and slope, potential changes in the "Level" attribute corresponding to different age groups can be inferred.

Insights into Age-Related Dynamics: This visualization aids in understanding how the categorical variable "Level" varies with age, providing insights into age-related dynamics within the dataset.

```
# Load necessary library
library(ggplot2)

# Assuming your dataset is loaded or created with the name 'data'

# Create a scatter plot for Age against Level with identifiable data points
ggplot(data = data, aes(x = Age, y = Level)) +
  geom_point(color = "blue", size = 3, alpha = 0.6) +  # Adding points with specified color, size, and transparency
  labs(x = "Age", y = "Level") +  # Labeling the axes
  ggtitle("Scatter Plot of Age vs. Level") +  # Adding a title
  theme_minimal()  # Applying a minimalistic theme
```



Age vs. Level Relationship: The scatter plot visualizes the relationship between age and a categorical variable named "Level".

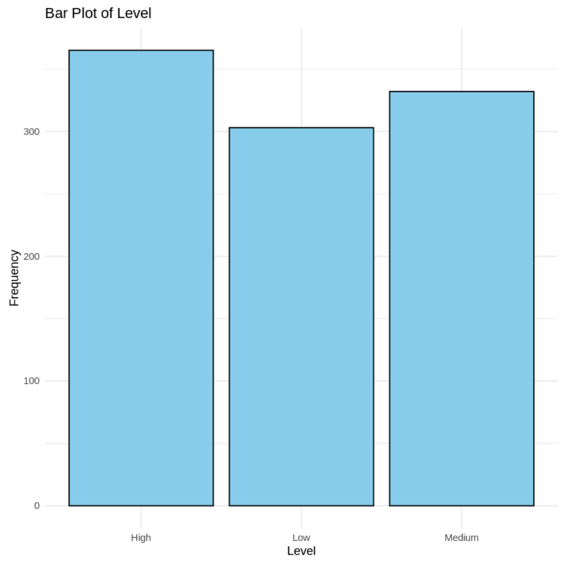
Individual Data Points: Each point on the plot represents an individual data point from the dataset, where the x-coordinate corresponds to the age of the individual and the y-coordinate corresponds to their level category.

Identifiable Data Points: The points on the scatter plot are identifiable due to their blue color, larger size (size = 3), and slight transparency (alpha = 0.6), making them easier to distinguish from each other and facilitating visual analysis.

Data Distribution: The scatter plot allows for the observation of how the "Level" variable is distributed across different age groups. The density or clustering of points at various age ranges provides insights into the prevalence or distribution of different levels within the dataset.

Insight into Age-Related Patterns: By examining the distribution and clustering of points, potential patterns or trends related to age and level categories can be inferred, aiding in understanding the relationship between age and the categorical variable "Level" within the dataset.

```
# Load necessary library
library(ggplot2)

# Assuming your dataset is loaded or created with the name 'data'

# Create a bar plot for the 'Level' variable
ggplot(data = data, aes(x = Level)) +
  geom_bar(fill = "skyblue", color = "black") +  # Adding bars with specified fill color and outline color
  labs(x = "Level", y = "Frequency", title = "Bar Plot of Level") +  # Labeling the axes and title
  theme_minimal()  # Applying a minimalistic theme
```



Level Distribution: The bar plot visualizes the distribution of the categorical variable "Level" within the dataset.

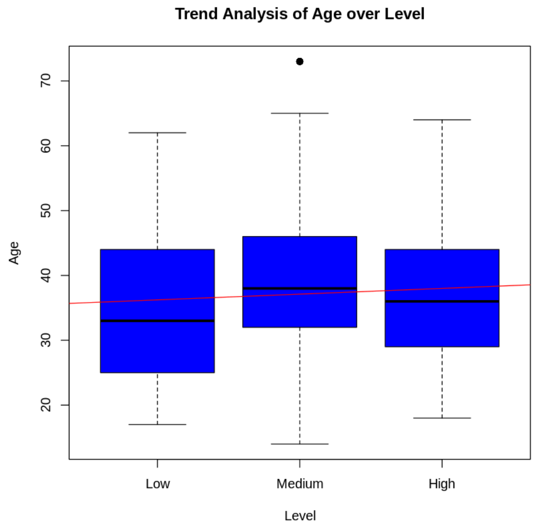
Bar Height: The height of each bar represents the frequency of data points belonging to each level category.

Color Specification: The bars of the plot are filled with a sky blue color (fill = "skyblue") and outlined in black (color = "black"), enhancing visual clarity and distinction.

Frequency Interpretation: The taller bars indicate higher frequencies of occurrence for particular level categories within the dataset.

Insights into Level Distribution: By examining the bar plot, insights into the distribution and prevalence of different level categories can be gained, facilitating the exploration and understanding of the data's characteristics.

```
# Read the dataset
data <- read.csv("carcinoma.csv")

# Convert 'Level' to a factor
data$Level <- factor(data$Level, levels = c("Low", "Medium", "High"))

# Trend analysis of Age over Level
plot(Age ~ Level, data = data,
     col = "blue",  # Set color of the points
     pch = 19,      # Set point type (filled circles)
     main = "Trend Analysis of Age over Level",  # Main title of the plot
     xlab = "Level",    # Label for x-axis
     ylab = "Age"       # Label for y-axis
)

# Add a trend line
abline(lm(Age ~ as.numeric(Level), data = data), col = "red")
```


Trend Analysis of Age over Level
Age
Low
Medium
High
Level