Data Analysis and Statistics (CSE-883) – Assignment 02

Submitted By

Muhammad Abdur Rehman

Faculty

Dr. Zamir Hussain

School of Interdisciplinary Engineering and Sciences, National University of Sciences and Technology, Islamabad

December, 2025

-data-analytics-and-statistics-

November 30, 2025

0.1 # Assignment 02 - Data Analytics and Statistics (CSE -883)

Submitted by:

Muhammad Abdur Rehman (538442)

Supervised By:

Dr. Zamir Hussain

---

a. Preprocess the file.

Note that the preprocessing includes dealing typos/errors/missing values/outliers, encoding and any other anomalies which you feel appropriate.

b. Consider the following objective:

To estimate the prevalence of cardiovascular risk factors (major risk factors, socio-demographic and behavioral determinants) among patients with diabetes mellitus.

Provide:

1) A most suitable simple linear regression model considering CRD Score as dependent variable.
2) An adequate multiple linear regression model considering CRD Score as dependent variable.

0.2 Data Analysis and Preprocessing

Load and inspect the dataset.

[37]:
```
import pandas as pd
data = pd.read_excel('/content/DAS-Assignment02-Data.xlsx')
data
```

[37]:

	education	Age	Gender	Smoking	BP	BMI	CVD Risk	WeightKg	\
0	Graduation	43	F	No	90 / 130	30.120482	2	83	
1	Graduation	45	M	No	80 / 120	29.372397	3	79	
2	No	42	F	No	90 / 150	37.182073	3	94	
3	Primary	51	F	No	80 / 120	34.232692	3	76	
4	No	42	F	No	70 / 120	26.106562	2	66	
…	…	…	…	…	…	…	…	…	
1300	Graduation	55	F	No	80 / 110	33.320518	4	78	

```
1301  Graduation   42      F      No  80 / 130  30.062102        2        76
1302          No   40      F      No  90 / 140  28.250970        2        67
1303          No   59      F      No  90 / 150  28.353057        7        69
1304     Masters   63      M      No  70 / 150  24.538965       11        66

      HeightCm  Height 2
0          166     27556
1          164     26896
2          159     25281
3          149     22201
4          159     25281
…          …       …
1300       153     23409
1301       159     25281
1302       154     23716
1303       156     24336
1304       164     26896

[1305 rows x 10 columns]
```

[38]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1305 entries, 0 to 1304
Data columns (total 10 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   education  1305 non-null   object
 1   Age        1305 non-null   int64
 2   Gender     1305 non-null   object
 3   Smoking    1305 non-null   object
 4   BP         1305 non-null   object
 5   BMI        1305 non-null   float64
 6   CVD Risk   1299 non-null   object
 7   WeightKg   1305 non-null   int64
 8   HeightCm   1305 non-null   int64
 9   Height 2   1305 non-null   int64
dtypes: float64(1), int64(4), object(5)
memory usage: 102.1+ KB
```

According to the data info, the dataset contains 5 columns, 'education', 'Gender', 'Smoking', 'BP', and 'CVD Risk' which contain categorical values (object), and multiple null values in 'CVD Risk' column.

Data description tells us about the basic statistics for every column present in the dataframe. Note that while describing the data, the categorical columns have been removed automatically becuase they containn Nan (Not a number) values.

[39]:
```
data.describe()
```

[39]:

	Age	BMI	WeightKg	HeightCm	Height 2
count	1305.000000	1305.000000	1305.000000	1305.000000	1305.000000
mean	56.199234	29.181310	76.075096	162.158621	26395.757088
std	8.448064	9.736734	14.134939	10.020767	3202.097278
min	40.000000	13.595895	31.000000	72.000000	5184.000000
25%	50.000000	25.333333	66.000000	155.000000	24025.000000
50%	57.000000	28.393726	75.000000	162.000000	26244.000000
75%	63.000000	31.957633	85.000000	170.000000	28900.000000
max	74.000000	320.216049	166.000000	190.000000	36100.000000

0.2.1 Encoding of Categorical Variables

Mapping of ordinal categories The education column was first standardized by removing extra spaces and converting all entries to uppercase to ensure consistency. A custom mapping was then applied to convert the ordinal categories into numeric values, reflecting the natural order of education levels from 0 (No/Not Done) to 5 (M.Phil/PhD).

[40]:
```
print("Original unique values in 'education' column:")
display(data['education'].unique())
```

```
Original unique values in 'education' column:
```

```
array(['Graduation', 'No', 'Primary', 'Secondary', 'NOT DONE', 'Masters',
       'M.Phil/Phd'], dtype=object)
```

[41]:
```
# Remove spaces and convert to consistent case
data['education'] = data['education'].str.strip().str.upper()

# Update mapping keys to match standardized strings
education_mapping = {
    'NO': 0,
    'NOT DONE': 0,
    'PRIMARY': 1,
    'SECONDARY': 2,
    'GRADUATION': 3,
    'MASTERS': 4,
    'M.PHIL/PHD': 5
}

# Map the values
data['education'] = data['education'].map(education_mapping)
```

[42]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1305 entries, 0 to 1304
```

```
Data columns (total 10 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   education   1305 non-null   int64
 1   Age         1305 non-null   int64
 2   Gender      1305 non-null   object
 3   Smoking     1305 non-null   object
 4   BP          1305 non-null   object
 5   BMI         1305 non-null   float64
 6   CVD Risk    1299 non-null   object
 7   WeightKg    1305 non-null   int64
 8   HeightCm    1305 non-null   int64
 9   Height 2    1305 non-null   int64
dtypes: float64(1), int64(5), object(4)
memory usage: 102.1+ KB
```

Label encoding for nominal categories For the nominal columns 'Gender' and 'Smoking', LabelEncoder was used to transform the categorical values into binary labels, as these variables have no inherent order and must be represented numerically for machine learning models.

[43]:
```
print("Original unique values in 'Smoking' column:")
display(data['Smoking'].unique())
```

```
Original unique values in 'Smoking' column:

array(['No', 'Cigarettes', 'Tobacco', 'Cigar'], dtype=object)
```

Males (M) were assigned 0 label, while females were assigned 1.

For Smoking column, non-smokers (No) were assigned 0, while all other categories of smokers (Cigarettes, Tobacco, Cigar) were assigned 1.

[44]:
```
data['Gender'] = data['Gender'].map({'M': 0, 'F': 1})

# Map Smoking: No = 0, everything else = 1
data['Smoking'] = data['Smoking'].apply(lambda x: 0 if x == 'No' else 1)

data
```

[44]:

	education	Age	Gender	Smoking	BP	BMI	CVD Risk	WeightKg	\
0	3	43	1	0	90 / 130	30.120482	2	83	
1	3	45	0	0	80 / 120	29.372397	3	79	
2	0	42	1	0	90 / 150	37.182073	3	94	
3	1	51	1	0	80 / 120	34.232692	3	76	
4	0	42	1	0	70 / 120	26.106562	2	66	
…	…	…	…	…	…	…	…	…	
1300	3	55	1	0	80 / 110	33.320518	4	78	
1301	3	42	1	0	80 / 130	30.062102	2	76	
1302	0	40	1	0	90 / 140	28.250970	2	67	

```
1303          0   59       1        0  90 / 150  28.353057        7        69
1304          4   63       0        0  70 / 150  24.538965       11        66

      HeightCm  Height 2
0          166     27556
1          164     26896
2          159     25281
3          149     22201
4          159     25281
…          …       …
1300       153     23409
1301       159     25281
1302       154     23716
1303       156     24336
1304       164     26896

[1305 rows x 10 columns]
```

0.2.2 Processing of the BP column

The blood pressure (BP) column was split into two separate columns, systolic and diastolic, by dividing the values at the "/" character. Extra spaces were removed, and the resulting values were converted to numeric format to allow quantitative analysis.

The original BP column was then dropped, leaving two clean numeric columns that can be used directly as features in the regression model.

[45]:
```
# Split BP into two new columns
data[['diastolic', 'systolic']] = data['BP '].str.split('/', expand=True)

# Remove extra spaces and convert to numeric
data['diastolic'] = data['diastolic'].str.strip().astype(float)
data['systolic'] = data['systolic'].str.strip().astype(float)

# Drop the original column
data = data.drop(columns=['BP '])
data
```

[45]:
```
      education  Age  Gender  Smoking       BMI  CVD Risk  WeightKg  HeightCm  \
0             3   43       1        0  30.120482         2        83       166
1             3   45       0        0  29.372397         3        79       164
2             0   42       1        0  37.182073         3        94       159
3             1   51       1        0  34.232692         3        76       149
4             0   42       1        0  26.106562         2        66       159
…           …  …     …       …        …       …         …        …
1300          3   55       1        0  33.320518         4        78       153
1301          3   42       1        0  30.062102         2        76       159
1302          0   40       1        0  28.250970         2        67       154
```

```
1303          0   59       1        0  28.353057        7        69       156
1304          4   63       0        0  24.538965       11        66       164

      Height 2  diastolic  systolic
0        27556       90.0     130.0
1        26896       80.0     120.0
2        25281       90.0     150.0
3        22201       80.0     120.0
4        25281       70.0     120.0
…          …          …        …
1300     23409       80.0     110.0
1301     25281       80.0     130.0
1302     23716       90.0     140.0
1303     24336       90.0     150.0
1304     26896       70.0     150.0

[1305 rows x 11 columns]
```

[46]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1305 entries, 0 to 1304
Data columns (total 11 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   education  1305 non-null   int64
 1   Age        1305 non-null   int64
 2   Gender     1305 non-null   int64
 3   Smoking    1305 non-null   int64
 4   BMI        1305 non-null   float64
 5   CVD Risk   1299 non-null   object
 6   WeightKg   1305 non-null   int64
 7   HeightCm   1305 non-null   int64
 8   Height 2   1305 non-null   int64
 9   diastolic  1305 non-null   float64
 10  systolic   1305 non-null   float64
dtypes: float64(3), int64(7), object(1)
memory usage: 112.3+ KB
```

0.2.3 Processing of the CVD Risk column

The CVD Risk column contained numeric values representing cardiovascular risk scores, along with two special cases: missing values (NaN) and entries reported as ">30", indicating very high risk beyond the measurable scale.

In clinical practice, a score above 30% is considered "high risk" for developing cardiovascular disease within the next 10 years.

To handle this, the ">30" entries were replaced with 32 to reflect very high-risk individuals while

maintaining a discrete numeric target for regression.

The missing values were imputed using the median of the column which is an approach that preserves the overall distribution without being skewed by extreme values.

[47]:
```
print("Original unique values and their frequencies in 'CVD Risk' column:")
display(data['CVD Risk'].value_counts(dropna=False))
```

```
Original unique values and their frequencies in 'CVD Risk' column:

CVD Risk
6      154
3      142
5      142
4      131
7      108
8       93
2       82
9       76
10      68
11      62
12      56
14      33
13      33
15      29
16      22
19      15
17      14
18      13
NaN      6
21       6
1        4
20       4
22       4
24       3
>30      2
23       1
25       1
28       1
Name: count, dtype: int64
```

[48]:
```
data['CVD Risk'] = data['CVD Risk'].replace('>30', 32)

median_value = data['CVD Risk'].median()
data['CVD Risk'] = data['CVD Risk'].fillna(median_value)
```

```
/tmp/ipython-input-1107437140.py:1: FutureWarning: Downcasting behavior in
`replace` is deprecated and will be removed in a future version. To retain the
old behavior, explicitly call `result.infer_objects(copy=False)`. To opt-in to
```

```
the future behavior, set `pd.set_option('future.no_silent_downcasting', True)`
  data['CVD Risk'] = data['CVD Risk'].replace('>30', 32)
```

[49]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1305 entries, 0 to 1304
Data columns (total 11 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   education  1305 non-null   int64
 1   Age        1305 non-null   int64
 2   Gender     1305 non-null   int64
 3   Smoking    1305 non-null   int64
 4   BMI        1305 non-null   float64
 5   CVD Risk   1305 non-null   float64
 6   WeightKg   1305 non-null   int64
 7   HeightCm   1305 non-null   int64
 8   Height 2   1305 non-null   int64
 9   diastolic  1305 non-null   float64
 10  systolic   1305 non-null   float64
dtypes: float64(4), int64(7)
memory usage: 112.3 KB
```

[50]:
```
data
```

[50]:

	education	Age	Gender	Smoking	BMI	CVD Risk	WeightKg	HeightCm	Height 2	diastolic	systolic
0	3	43	1	0	30.120482	2.0	83	166	27556	90.0	130.0
1	3	45	0	0	29.372397	3.0	79	164	26896	80.0	120.0
2	0	42	1	0	37.182073	3.0	94	159	25281	90.0	150.0
3	1	51	1	0	34.232692	3.0	76	149	22201	80.0	120.0
4	0	42	1	0	26.106562	2.0	66	159	25281	70.0	120.0
…	…	…	…	…	…	…	…	…	…	…	…
1300	3	55	1	0	33.320518	4.0	78	153	23409	80.0	110.0
1301	3	42	1	0	30.062102	2.0	76				
1302	0	40	1	0	28.250970	2.0	67				
1303	0	59	1	0	28.353057	7.0	69				
1304	4	63	0	0	24.538965	11.0	66				

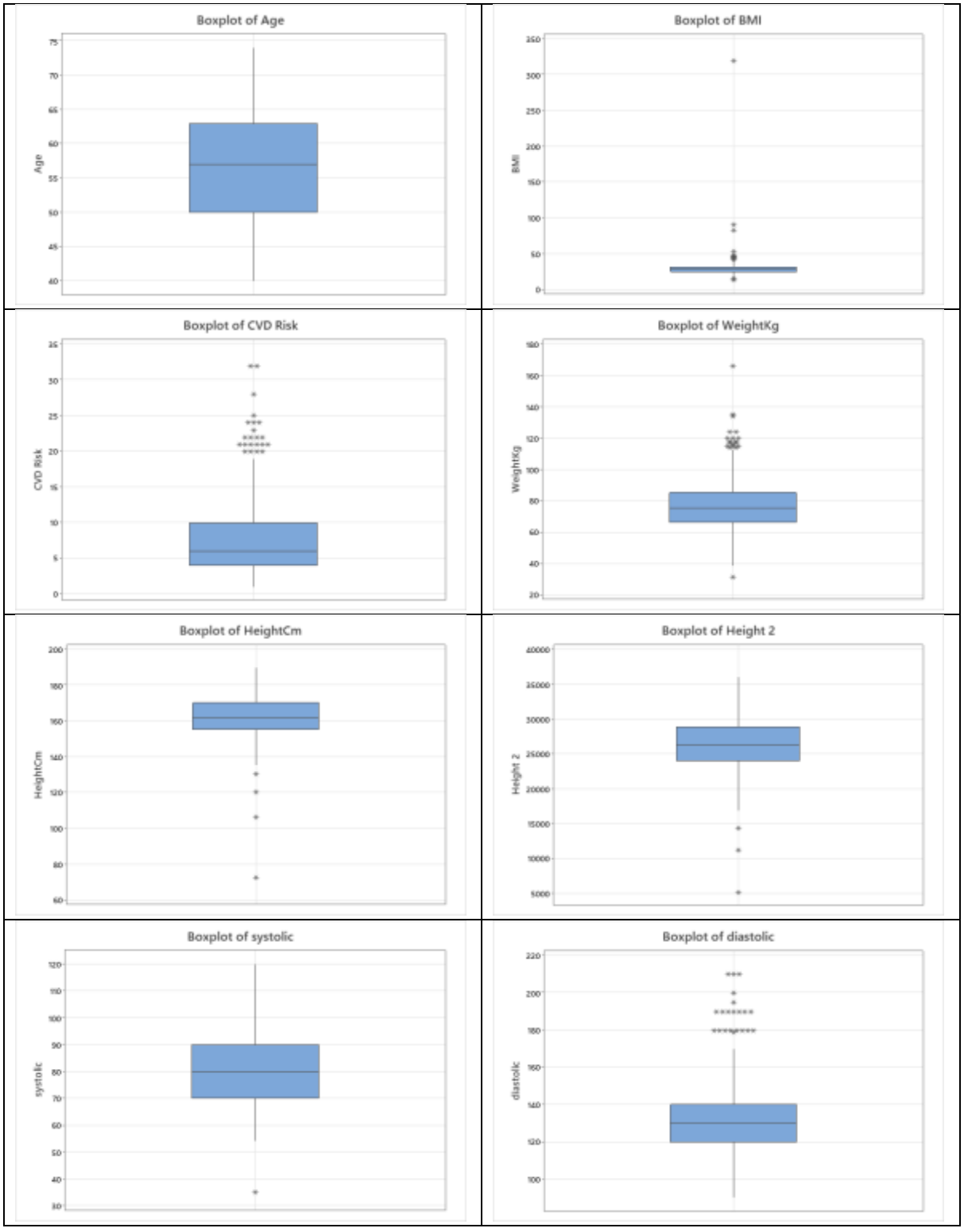
```
1301       159     25281       80.0     130.0
1302       154     23716       90.0     140.0
1303       156     24336       90.0     150.0
1304       164     26896       70.0     150.0

[1305 rows x 11 columns]
```

[51]:
```
df = pd.DataFrame(data)
```

[52]:
```
df.to_excel("data before outlier removal.xlsx", index=False)
```

Boxplots before cleaning (outlier removal):

We generate boxplots to visually identify outliers present in the dataset.



Descriptive statistics before outlier removal:

Descriptive analysis for the dataset was performed before outlier handling using Minitab.

Variable	Mean	SE.Mean	St. Dev	Variance	Minimum	Q1	Median	Q3	Maximum
Age	56.19923	0.233858	8.448064	71.36	40	50	57	63	74
BMI	29.18131	0.269531	9.736734	94.804	13.59589	25.33333	28.39373	31.95763	320.216
CVD Risk	7.622222	0.125341	4.527926	20.50	1	4	6	10	32
WeightKg	76.0751	0.391281	14.13494	199.79	31	66	75	85	166
HeightCm	162.1586	0.277393	10.02077	100.41	72	155	162	170	190
Height 2	26395.76	88.6399	3202.097	10253427	5184	24025	26244	28900	36100
Diastolic	80.23065	0.268008	9.681729	93.73	35	70	80	90	120
Systolic	134.3234	0.453936	16.39836	268.90	90	120	130	140	210

Variable	Range	IQR	Skewness	Kurtosis
Age	34	13	-0.00231	-0.87734
BMI	306.6202	6.6243	20.87142	611.8152
CVD Risk	31	6	1.229162	1.932592
WeightKg	135	19	0.635321	1.620082
HeightCm	118	15	-0.58168	5.050138
Height 2	30916	4875	-0.0576	1.26567
Diastolic	85	20	0.08305	0.378658
Systolic	120	20	0.467292	1.350565

Looking at these preliminary statistics, the age distribution looks reasonable, centered around 56 years, with most participants falling between 50 and 63.

While the BMI median sits at a fairly typical 28.4, there's an extreme maximum value of 320 that's clearly erroneous, because no human could survive with that BMI. This is pulling the mean up and creating massive skewness. The interquartile range suggests most people have relatively normal BMIs, but that outlier is distorting everything.

Cardiovascular risk scores range from 1 to 32, with half the sample scoring between 4 and 10. The distribution skews right, meaning more people cluster at lower risk levels with a tail extending toward higher risk.

Weight and height measurements appear mostly sensible. Weight averages 76 kg, with most people between 66 and 85 kg, though someone is also recorded at 31 kg. Heights center around 162 cm, which suggests either a predominantly female sample or participants from populations with shorter average stature.

Overall, the data shows signs of outliers that need cleaning before any real analysis can begin. The extreme BMI values are the most pressing issues to address. Overall, this appears to be a cardiovascular health study of an older population with some outliers that need cleaning.

0.1 Outlier Removal

To identify outliers in the dataset, we applied the **Interquartile Range (IQR) method** to all numerical columns, as outlier logic doesn't apply to categorical variables.

First, the code calculated the **25th percentile (Q1), 75th percentile (Q3),** and the **IQR (Q3–Q1)** for each variable and stored these values in a dictionary. These statistics summarize the central spread of each feature.

Next, using the standard IQR rule, we computed the lower and upper cutoff points for outlier detection:

Q1 − 1.5×IQR

Q3 + 1.5×IQR

Any data point falling outside this range was flagged as a potential outlier.

A separate DataFrame **(outlier_flags)** was then created to store a True/False label for every value, indicating whether it was an outlier.

Finally, the script counted how many outliers appeared in each column by summing the True values. This gave a clear overview of which variables contained extreme values and how many observations fell outside the expected range.

[ ]:
```
# Drop categorical columns for outlier analysis
numerical_columns = data.drop(columns=['education','Smoking', 'Gender'])

# Display included columns
print("Columns included in outlier detection:")
print(list(numerical_columns.columns))
```

```
Columns included in outlier detection:
['Age', 'BMI ', 'CVD Risk', 'WeightKg', 'HeightCm', 'Height 2', 'diastolic',
'systolic']
```

0.1.1 Outlier detection

[ ]:
```
# Create a dictionary to store Q1, Q3, and IQR values
iqr_values = {}

for col in numerical_columns.columns:
```

```
    Q1 = numerical_columns[col].quantile(0.25)
    Q3 = numerical_columns[col].quantile(0.75)
    IQR = Q3 - Q1
    iqr_values[col] = (Q1, Q3, IQR)

# Create an outlier flag DataFrame
outlier_flags = pd.DataFrame(index=numerical_columns.index)

for col, (Q1, Q3, IQR) in iqr_values.items():
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    outlier_flags[col] = (numerical_columns[col] < lower_bound) |␣
 ↪(numerical_columns[col] > upper_bound)

# Count outliers per column
print("\nOutliers detected per column:")
print(outlier_flags.sum())
```

```
Outliers detected per column:
Age           0
BMI          24
CVD Risk     22
WeightKg     19
HeightCm      4
Height 2      3
diastolic     1
systolic     21
dtype: int64
```

For moderate datasets, removing extreme outliers is usually fine, especially if they are a tiny fraction of total rows.

[ ]:
```
# Identify rows containing at least one outlier
rows_with_outliers = outlier_flags.any(axis=1)

# Display summary
print(f"\nTotal rows with outliers: {rows_with_outliers.sum()}")

# Remove outlier rows
data_cleaned = data[~rows_with_outliers].reset_index(drop=True)

print(f"Shape before cleaning: {data.shape}")
print(f"Shape after cleaning: {data_cleaned.shape}")

# Preview cleaned dataset
data_cleaned.head()
```

```
Total rows with outliers: 70
Shape before cleaning: (1305, 11)
Shape after cleaning: (1235, 11)
```

[ ]:
```
   education  Age  Gender  Smoking       BMI   CVD Risk  WeightKg  HeightCm  \
0          3   43       1        0  30.120482       2.0        83       166
1          3   45       0        0  29.372397       3.0        79       164
2          0   42       1        0  37.182073       3.0        94       159
3          1   51       1        0  34.232692       3.0        76       149
4          0   42       1        0  26.106562       2.0        66       159

   Height 2  diastolic  systolic
0     27556       90.0     130.0
1     26896       80.0     120.0
2     25281       90.0     150.0
3     22201       80.0     120.0
4     25281       70.0     120.0
```

0.2 Final dataset

After these preprocessing steps, the dataset is clinically interpretable and suitable for use as the input in the regression model.

[ ]:
```
data_cleaned
```

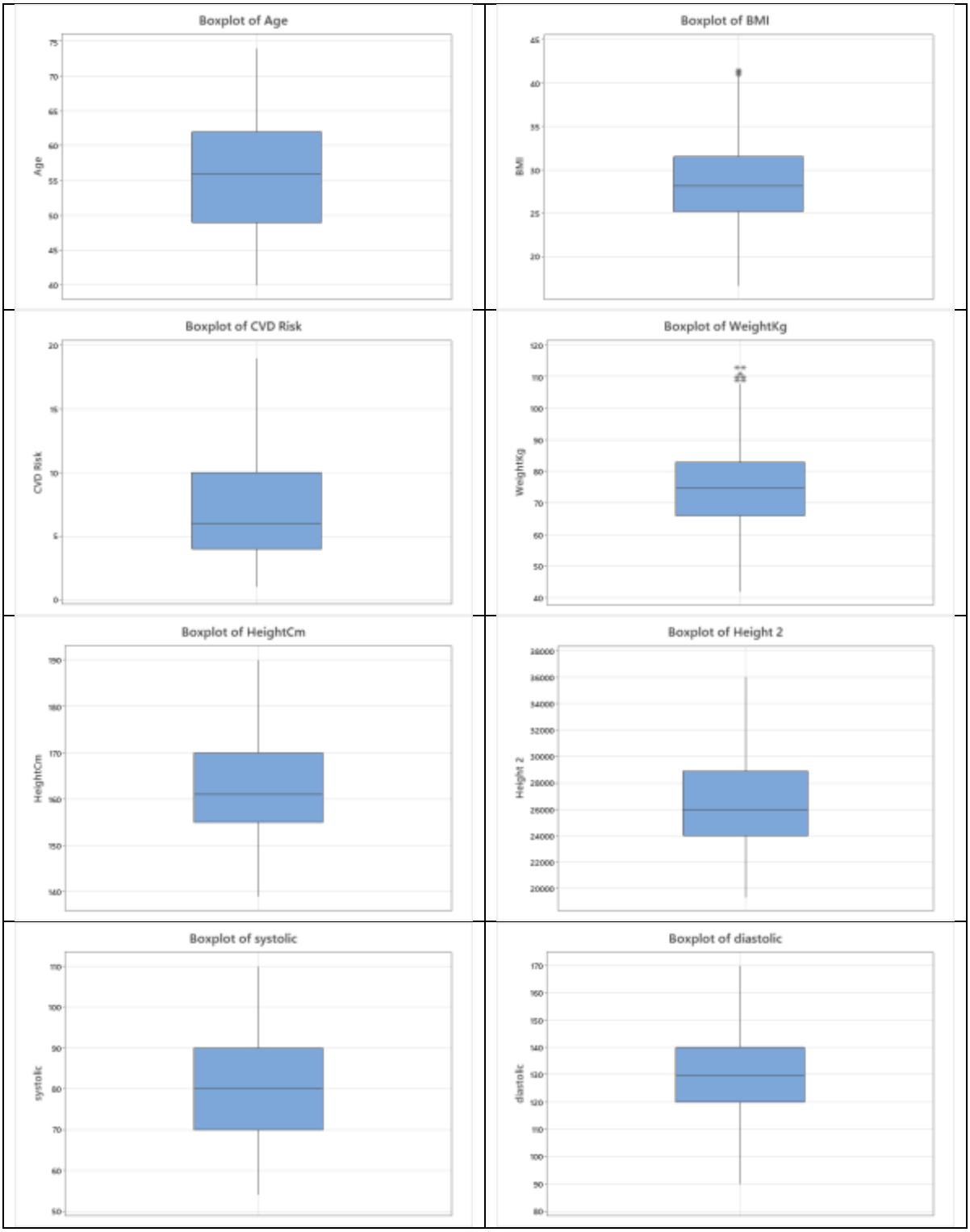
[ ]:
```
      education  Age  Gender  Smoking       BMI   CVD Risk  WeightKg  \
0             3   43       1        0  30.120482       2.0        83
1             3   45       0        0  29.372397       3.0        79
2             0   42       1        0  37.182073       3.0        94
3             1   51       1        0  34.232692       3.0        76
4             0   42       1        0  26.106562       2.0        66
...         ...  ...     ...      ...        ...       ...
1230          3   55       1        0  33.320518       4.0        78
1231          3   42       1        0  30.062102       2.0        76
1232          0   40       1        0  28.250970       2.0        67
1233          0   59       1        0  28.353057       7.0        69
1234          4   63       0        0  24.538965      11.0        66

      HeightCm  Height 2  diastolic  systolic
0          166     27556       90.0     130.0
1          164     26896       80.0     120.0
2          159     25281       90.0     150.0
3          149     22201       80.0     120.0
4          159     25281       70.0     120.0
...        ...       ...        ...       ...
1230       153     23409       80.0     110.0
1231       159     25281       80.0     130.0
```

```
1232       154     23716       90.0     140.0
1233       156     24336       90.0     150.0
1234       164     26896       70.0     150.0

[1235 rows x 11 columns]
```

[ ]:
```
data_cleaned.to_excel('cleaned_data.xlsx', index=False)
print("DataFrame saved to cleaned_data.xlsx.")
```

```
DataFrame saved to cleaned_data.xlsx.
```

Boxplots after outlier removal:



Descriptive statistics after outlier removal:

Variable	Mean	SE.Mean	St. Dev	Variance	Minimum	Q1	Median	Q3	Maximum
Age	56.0170	0.238647	8.38666	70.3360	40	49	56	62	74
BMI	28.5429	0.128184	4.50473	20.2926	16.6133	25.2174	28.1625	31.5334	41.4738
CVD Risk	7.26802	0.114707	4.03111	16.2498	1	4	6	10	19
WeightKg	75.0980	0.360879	12.6822	160.839	42	66	75	83	113
HeightCm	162.304	0.266505	9.36566	87.7157	139	155	161	170	190
Height 2	26430.1	87.0882	3060.50	9366673	19321	24025	25921	28900	36100
diastolic	79.8154	0.259970	9.13601	83.4667	54	70	80	90	110
systolic	133.189	0.418515	14.7077	216.316	90	120	130	140	170

Variable	Range	IQR	Skewness	Kurtosis
Age	34	13	0.00	-0.88
BMI	24.8605	6.31609	0.35	-0.27
CVD Risk	18	6	0.87	0.16
WeightKg	71	17	0.20	-0.24
HeightCm	51	15	0.21	-0.57
Height 2	16779	4875	0.33	-0.47
diastolic	56	20	-0.06	-0.38
systolic	80	20	-0.11	-0.27

Here are the 5 key changes after outlier removal:

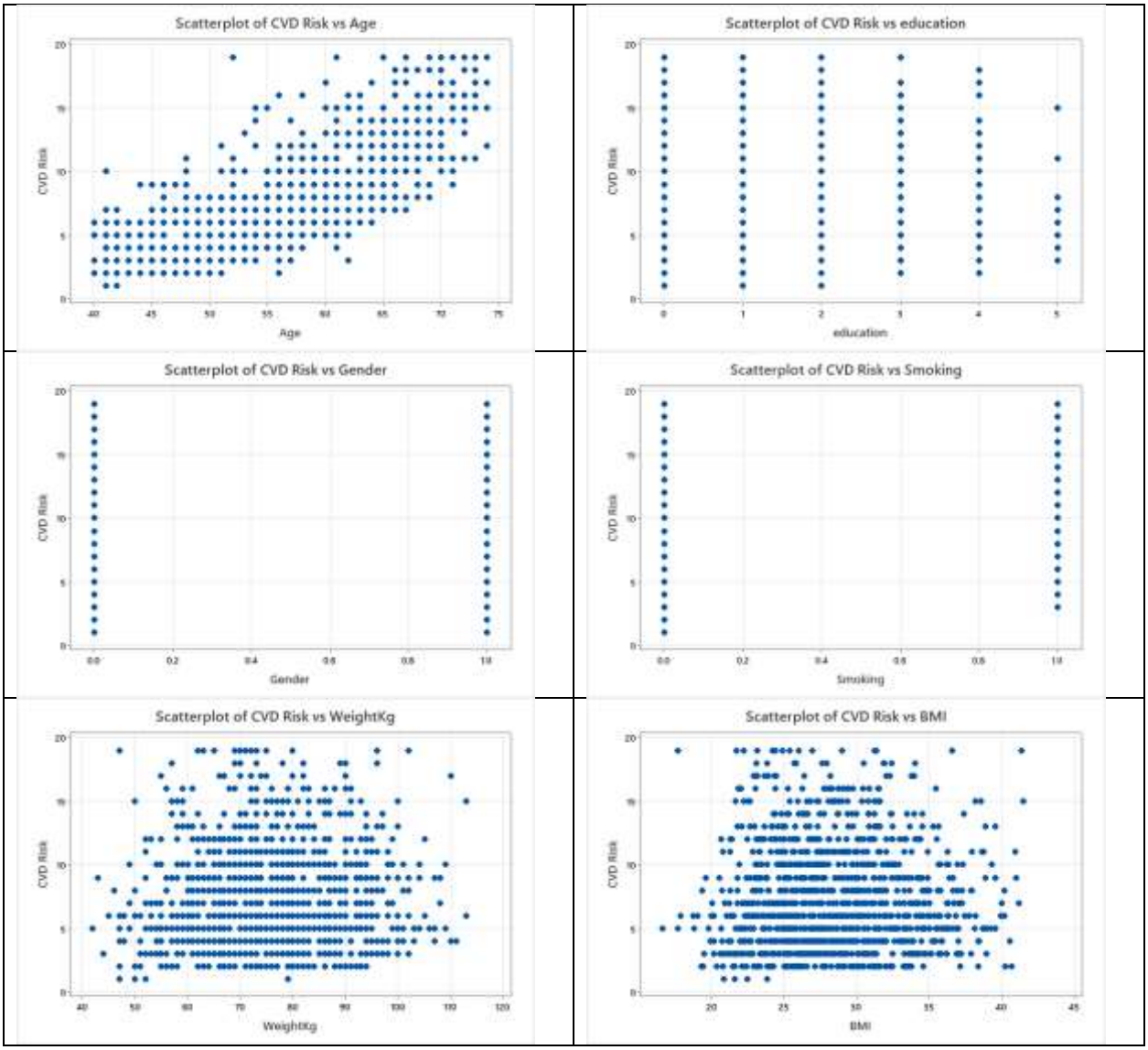
1. **BMI:** The maximum of 320 dropped to a realistic 41.5, and the standard deviation reduced from 9.7 to 4.5. The skewness went from 20.87 down to 0.35, meaning the distribution is now nearly symmetrical instead of being distorted.
2. **Weight range:** Maximum weight fell from 166 kg to 113 kg, and more importantly the 31 kg minimum is gone (now 42 kg). The distribution went from positively skewed (0.64) to almost perfectly normal (0.20).
3. **Blood pressure extremes were trimmed**: Diastolic lost that 35 reading (now starts at 54) and systolic's 210 peak came down to 170. Both distributions moved from slightly skewed to essentially normal, which makes more physiological sense.
4. **CVD risk became more moderate**: The maximum score dropped from 32 to 19, cutting off the extreme high-risk tail. Skewness decreased from 1.23 to 0.87, though there's still a lean toward lower risk scores in the sample.

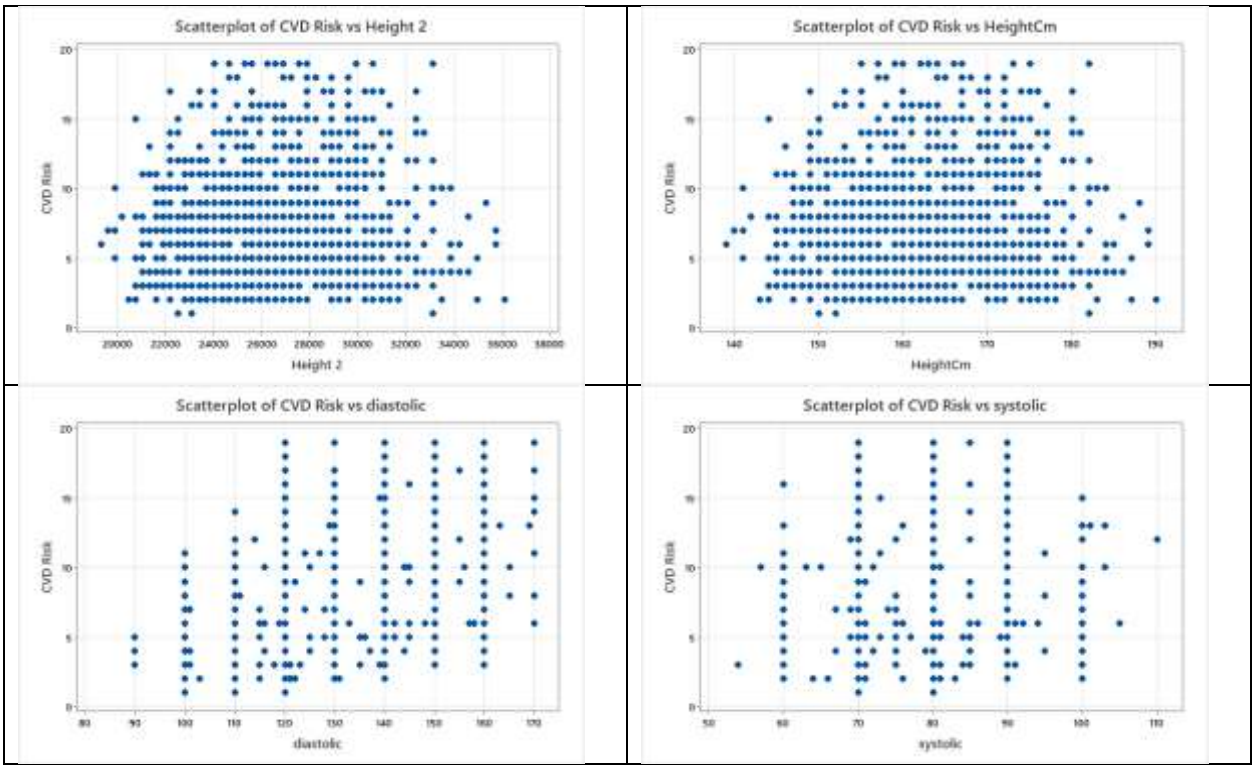
Correlation Analysis:

Correlation analysis is essential before fitting a simple linear regression because it helps to verify if there's actually a meaningful linear relationship between the variables. There's no point building a regression model if the correlation is weak or the relationship isn't linear to begin with.

Since our dependent variable is "CVD Risk", we will perform a correlation analysis and choose the independent variable with the highest correlation with the target to develop a simple linear regression model.

Before doing that, we will visually identify the relationship between IDVs and DV using a scatter plot.





The scatter plots indicate that "Age" is the only variable with a linear positive correlation with the dependent variable (CVD Risk).

Pearson's Correlations:

	education	Age	Gender	Smoking	BMI	CVD Risk	Weight Kg	HeighCm	Height 2	diastolic
Age	-0.120									
Gender	-0.359	0.020								
Smoking	0.083	-0.052	-0.326							
BMI	-0.081	-0.076	0.246	-0.153						
CVDRisk	-0.017	0.796	-0.255	0.238	-0.004					
WeightKg	0.160	-0.133	-0.282	0.039	0.748	0.083				
HeightCm	0.342	-0.088	-0.747	0.273	-0.248	0.126	0.450			
Height 2	0.341	-0.089	-0.743	0.272	-0.248	0.123	0.450	0.999		
Diastolic	-0.035	-0.096	-0.068	-0.040	0.147	0.117	0.174	0.046	0.048	
Systolic	-0.093	0.102	0.074	-0.125	0.169	0.348	0.108	-0.079	-0.078	0.618

From the correlation analysis, it is evident that the independent variable, "Age," exhibits the highest correlation with the target (+0.796), which is positive, strong, and statistically significant (p-value < 0.05, as shown below). Therefore, to build our simple linear regression model, we will choose this variable.

Regression equation:

CVD Risk = -14.158 + 0.38249 Age

Model summary:

S	R-sq	R-sq(adj)	R-sq(pred)
2.44222	63.33%	63.30%	63.20%

Analysis of variance:

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	12698.2	12698.2	2128.98	0.000
Age	1	12698.2	12698.2	2128.98	0.000
Error	1233	7354.1	6.0		
Lack-of-Fit	33	846.8	25.7	4.73	0.000
Pure Error	1200	6507.3	5.4		
Total	1234	20052.3			

Interpretation:

Looking at this regression model, age is a statistically significant predictor of CVD risk (p-value of 0.00 means we can be extremely confident this relationship isn't due to chance).

The equation tells us that for every additional year of age, the CVD risk score increases by about 0.38 points on average. So if you compare someone who's 60 to someone who's 50, you'd expect the older person to have a CVD risk score that's roughly 3.8 points higher (10 years × 0.38).

The intercept of -14.158 doesn't have a practical meaning here since it would represent the CVD risk at age zero, which is outside our data range and gives a negative risk score (impossible in reality). It's just the mathematical anchor point for the line.

The R-squared of 63.33% is pretty solid; it means age alone explains about 63% of the variation in CVD risk scores across individuals. That's actually quite good for a single predictor, though it also tells us there's still 37% of the variation that's driven by other factors not captured in this model (things like smoking, height, weight, etc.).

Conclusion:

Age is a strong predictor of cardiovascular risk in the dataset, but there's clearly room to improve the model by adding other relevant variables.

Multiple Linear Regression Model:

We previously discussed that if variables have no linear relationship with the target, we omit those variables from a simple linear regression model. However, for a multiple regression model, we might still want to include those variables because they might have important relationships with

the target variable that aren't captured by simple linear correlation. For example, they could have non-linear effects, interaction effects with other predictors, or suppressor effects where their contribution only becomes apparent when removing other variables in the model.

Backward elimination is a procedure in multiple regression model development where you start building a model including all the variables, and based on p-value and VIF, you sequentially remove the IDVs until the model becomes stable (there is no huge loss in the R-sq value). If a predictor is redundant because of correlation, it will automatically get a high p-value and be removed. At the end, you are left with an adequate set of IDVs.

1. **Model including all variables:**

We started building the model with all the variables. The regression equations for all values of categorical predictors are as follows:

Note:
Continuous variable equation (CVE) = + 0.37963 Age - 0.1642 BMI + 0.0994 WeightKg +0.162 HeightCm - 0.000776 Height 2 - 0.00993 diastolic + 0.08548 systolic

education	Gender	Smoking	Regression equation
0	0	0	CVD Risk = **-32.7** + CVE
0	0	1	CVD Risk = **-29.5** + CVE
0	1	0	CVD Risk = **-34.5** + CVE
0	1	1	CVD Risk = **-31.4** + CVE
1	0	0	CVD Risk = **-32.7** + CVE
1	0	1	CVD Risk = **-29.5** + CVE
1	1	0	CVD Risk = **-34.5** + CVE
1	1	1	CVD Risk = **-31.4** + CVE
2	0	0	CVD Risk = **-32.4** + CVE
2	0	1	CVD Risk = **-29.3** + CVE
2	1	0	CVD Risk = **-34.3** + CVE
2	1	1	CVD Risk = **-31.2** + CVE
3	0	0	CVD Risk = **-32.6** + CVE
3	0	1	CVD Risk = **-29.5** + CVE
3	1	0	CVD Risk = **-34.5** + CVE
3	1	1	CVD Risk = **-31.3** + CVE
4	0	0	CVD Risk = **-32.6** + CVE
4	0	1	CVD Risk = **-29.4** + CVE
4	1	0	CVD Risk = **-34.5** + CVE
4	1	1	CVD Risk = **-31.3** + CVE
5	0	0	CVD Risk = **-33.0** + CVE
5	0	1	CVD Risk = **-29.8** + CVE
5	1	0	CVD Risk = **-34.9** + CVE
5	1	1	CVD Risk = **-31.7** + CVE

Model summary:

S	R-sq	R-sq(adj)	R-sq(pred)

1.53622	85.64%	85.48%	85.25%

Analysis of variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	14	17173.1	1226.7	519.77	0.000
Age	1	11396.1	11396.1	4828.93	0.000
BMI	1	7.7	7.7	3.24	0.072
WeightKg	1	18.8	18.8	7.98	0.005
HeightCm	1	2.8	2.8	1.20	0.274
Height 2	1	7.6	7.6	3.20	0.074
Diastolic	1	5.8	5.8	2.46	0.117
Systolic	1	1112.7	1112.7	471.50	0.000
education	5	11.6	2.3	0.98	0.426
Gender	1	416.4	416.4	176.42	0.000
Smoking	1	1156.5	1156.5	490.06	0.000
Error	1220	2879.2	2.4		
Lack-of-Fit	1218	2879.2	2.4	*	*
Pure Error	2	0.0	0.0		
Total	1234	20052.3			

Interpretation:

This regression model shows how CVD risk varies across different combinations of education level, gender, and smoking status, while keeping all the continuous predictors (Age, BMI, Weight, Height, blood pressure) with the same coefficients across all groups.

The continuous predictors work the same way for everyone:

- **Age** increases CVD risk by 0.38 points per year, still the strongest predictor
- **BMI** has a negative coefficient (-0.1642), which might indicate suppression since weight is also in the model
- **Weight** adds 0.0994 points per kg
- **Height** and **Height²** have opposing small effects that essentially create a slight curve
- **Diastolic BP** slightly decreases risk (-0.00993)
- **Systolic BP** increases risk by 0.085 points per mmHg

Looking at the pattern of intercepts:

- **Smoking**: When smoking = 1, the intercept increases by about 3.1-3.2 points compared to smoking = 0. This suggests smokers start with a higher baseline CVD risk.
- **Gender**: When gender = 1 (females), the intercept drops by about 1.8-1.9 points compared to gender = 0 (males). This shows that males have a higher CVD risk

- **Education**: Higher education levels (4-5) have slightly lower intercepts than lower levels (0-1), suggesting that differences in education are minimal (less than 0.5 points across the range).

The intercepts range from about -29 to -35, indicating that the baseline shifts based on the demographic category, but once we account for that, age, blood pressure, and body metrics affect everyone the same way. The smoking effect appears strongest (3+ point difference), gender has a moderate effect (2 points), and education has minimal impact (less than 1 point across all levels).

With all the variables included, the R-sq value of the MLR model is 85.64%, which means that combining all the variables explains 85.64% of the variance in the CVD Risk, which is excellent for a health-related outcome.

However, the p-value of education is 0.426, which means that education is not a statistically significant variable to predict the change in 'CVD Risk', so in the next step, we will remove 'education'.

2. **After removing education:**

Regression equation:

Note:
Continuous variable equation (CVE) = + 0.37881 Age - 0.1619 BMI + 0.0991 WeightKg + 0.157 HeightCm - 0.000758 Height 2 - 0.01018 diastolic + 0.08557 systolic

Gender	Smoking	Regression equation
0	0	CVD Risk = **-32.3** + CVE
0	1	CVD Risk = **-29.2** + CVE
1	0	CVD Risk = **-34.2** + CVE
1	1	CVD Risk = **-31.1** + CVE

Model summary:

S	R-sq	R-sq(adj)	R-sq(pred)
1.53617	85.58%	85.48%	85.31%

Analysis of variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	9	17161.5	1906.8	808.04	0.000
Age	1	11532.3	11532.3	4886.96	0.000
BMI	1	7.5	7.5	3.16	0.076
WeightKg	1	18.7	18.7	7.94	0.005
HeightCm	1	2.7	2.7	1.14	0.286
Height 2	1	7.3	7.3	3.07	0.080
diastolic	1	6.1	6.1	2.59	0.108

systolic	1	1117.7	1117.7	473.63	0.000
Gender	1	443.5	443.5	187.96	0.000
Smoking	1	1165.5	1165.5	493.90	0.000
Error	1225	2890.8	2.4		
Lack-of-Fit	1222	2890.8	2.4	*	*
Pure Error	3	0.0	0.0		
Total	1234	20052.3			

Interpretation:

After removing education, the model simplified nicely and barely lost any predictive power (R-squared only dropped from 85.64% to 85.58%, which is just 0.06%). This confirms that education wasn't really contributing much to predicting CVD risk.

Males are 0, Females are 1:

Looking at the intercepts by group:

- **Males, non-smokers** (0,0): Baseline = -32.3
- **Males, smokers** (0,1): Baseline = -29.2 (about 3.1 points higher risk)
- **Females, non-smokers** (1,0): Baseline = -34.2 (about 1.9 points lower than males)
- **Females, smokers** (1,1): Baseline = -31.1 (combining both effects)

This makes biological sense!

The gender effect (~1.9 points): Males have a higher baseline CVD risk than females, which aligns perfectly with medical literature; men are at greater cardiovascular risk.

The smoking effect (~3.1 points): Smokers have about 3.1 points higher baseline CVD risk compared to non-smokers, regardless of gender. This is the stronger of the two categorical effects.

So the highest risk group would be male smokers (baseline -29.2), followed by male non-smokers (-32.3), then female smokers (-31.1), and the lowest baseline risk is female non-smokers (-34.2).

The model now tells a coherent story: being male and smoking both independently increase your cardiovascular risk, with smoking having a slightly larger impact than gender.

The p-values have been adjusted, and now 'HeightCm' has a p-value of 0.286, which means that this variable is not statistically significant in predicting. So, in the next step, we will eliminate 'HeightCm'.

3. **After removing HeightCm:**

Regression equation:

Note:
Continuous variable equation (CVE) = -18.67 + 0.37897 Age - 0.1874 BMI + 0.1091 WeightKg - 0.000308 Height 2 - 0.01067 diastolic + 0.08564 systolic

Gender	Smoking	Regression equation
0	0	CVD Risk = **-18.67** + CVE
0	1	CVD Risk = **-15.52** + CVE
1	0	CVD Risk = **-20.58** + CVE
1	1	CVD Risk = **-17.43** + CVE

Model summary:

S	R-sq	R-sq(adj)	R-sq(pred)
1.53626	85.57%	85.48%	85.32%

Analysis of variance:

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	8	17158.8	2144.9	908.80	0.000
Age	1	11551.1	11551.1	4894.34	0.000
BMI	1	10.7	10.7	4.55	0.033
WeightKg	1	24.4	24.4	10.34	0.001
Height 2	1	24.1	24.1	10.21	0.001
diastolic	1	6.8	6.8	2.87	0.091
systolic	1	1119.6	1119.6	474.39	0.000
Gender	1	473.4	473.4	200.59	0.000
Smoking	1	1166.0	1166.0	494.06	0.000
Error	1226	2893.5	2.4		
Lack-of-Fit	1223	2893.5	2.4	*	*
Pure Error	3	0.0	0.0		
Total	1234	20052.3			

Interpretation:

After ‘HeightCm’ was removed, the R-sq value dropped 0.01% which is not very significant. The p-values are adjusted, and now ‘diastolic’ has a p-value of 0.091, which means that this variable is not statistically significant to predict DV. So, in the next step, we will eliminate ‘diastolic’.

4. **After removing diastolic:**

Regression equation:

Note:
Continuous variable equation (CVE) = + 0.38073 Age - 0.1834 BMI + 0.1070 WeightKg - 0.000303 Height 2 + 0.08145 systolic

Gender	Smoking	Regression equation
0	0	CVD Risk = **-19.17** + CVE
0	1	CVD Risk = **-16.03** + CVE
1	0	CVD Risk = **-21.06** + CVE
1	1	CVD Risk = **-17.92** + CVE

5. **Model summary:**

S	R-sq	R-sq(adj)	R-sq(pred)
1.53742	85.54%	85.45%	85.31%

6. **Analysis of variance:**

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	7	17152.1	2450.3	1036.65	0.000
Age	1	12103.9	12103.9	5120.80	0.000
BMI	1	10.3	10.3	4.35	0.037
WeightKg	1	23.5	23.5	9.95	0.002
Height 2	1	23.3	23.3	9.85	0.002
systolic	1	1675.2	1675.2	708.74	0.000
Gender	1	466.8	466.8	197.51	0.000
Smoking	1	1165.9	1165.9	493.28	0.000
Error	1227	2900.2	2.4		
Lack-of-Fit	1221	2887.7	2.4	1.14	0.492
Pure Error	6	12.5	2.1		
Total	1234	20052.3			

Final multiple linear regression model interpretation

After conducting backward elimination based on p-values, the final model includes six predictors:

"Age, BMI, Weight, Height², systolic Blood Pressure, Gender, and Smoking Status"

The model demonstrates excellent overall fit with an R-squared of 85.54%, meaning it explains approximately 85.5% of the variation in CVD risk scores. The adjusted R-squared (85.45%) and predicted R-squared (85.31%) values are very close to the regular R-squared, indicating the model is not overfitted and should generalize well to new data. The residual standard error of 1.537 suggests predictions typically deviate from actual CVD risk scores by about 1.5 points.

Model significance

The ANOVA table reveals that the overall regression model is highly significant ($F = 1036.65$, $p < 0.000$), providing strong evidence that the predictors collectively have a meaningful relationship with CVD risk. All individual predictors in the final model are statistically significant, with p-values well below the conventional 0.05 threshold.

Age emerges as the strongest predictor ($F = 5120.80$, $p < 0.000$), with a coefficient of 0.381. This means each additional year of age increases CVD risk by approximately 0.38 points, holding all

other variables constant. Over a 10-year age difference, this translates to approximately 3.8 points in CVD risk, demonstrating the substantial impact of age on cardiovascular health.

Systolic blood pressure is the second most influential predictor ($F = 708.74$, $p < 0.000$), with a coefficient of 0.081. Each 1 mmHg increase in systolic pressure corresponds to about 0.08 points higher CVD risk. For someone with systolic BP 20 mmHg higher than average, this represents approximately 1.6 additional points of risk.

Smoking status shows a strong effect ($F = 493.28$, $p < 0.000$). Comparing the intercepts across smoking groups, smokers have approximately 3.1 points higher baseline CVD risk compared to non-smokers, regardless of gender. This substantial difference highlights smoking's cardiovascular hazards in medical literature.

Gender also plays a significant role ($F = 197.51$, $p < 0.000$). Males (coded as 0) have a baseline CVD risk approximately 1.9 points higher than females (coded as 1), consistent with epidemiological evidence that men face elevated cardiovascular risk compared to women.

Weight has a modest but significant positive effect ($F = 9.95$, $p = 0.002$), with a coefficient of 0.107. Each kilogram increase in body weight adds about 0.11 points to CVD risk. While this seems small per kilogram, a 20 kg weight difference translates to roughly 2 points of CVD risk.

BMI shows a negative coefficient of -0.183 ($F = 4.35$, $p = 0.037$), which appears counterintuitive at first glance. However, this likely reflects a **suppression effect** since weight is also in the model. When controlling for absolute weight, BMI's negative coefficient may indicate that for a given weight, having that weight distributed over a larger frame (higher height, thus lower BMI) is associated with slightly higher risk.

Height squared has a very small negative coefficient of -0.000303 ($F = 9.85$, $p = 0.002$). While statistically significant, its practical impact is minimal given the scale of the outcome variable. This variable was retained due to its statistical significance, but it contributes negligibly to predictions.

Categorical group comparisons

The model generates four distinct equations based on gender and smoking combinations. Male smokers exhibit the highest baseline CVD risk (intercept = -16.03), followed by female smokers (-17.92), male non-smokers (-19.17), and female non-smokers (-21.06) with the lowest baseline risk. The approximately 5-point spread between the highest and lowest risk groups highlights the combined impact of these demographic factors.

Conclusion

Through systematic backward elimination, variables with non-significant p-values (education, systolic blood pressure, and height) were removed, resulting in a parsimonious final model that retains excellent predictive performance. The final model achieves an R-squared of 85.54%, only

marginally lower than the full model's 85.64%, demonstrating that the removed variables contributed minimal explanatory power.

This analysis successfully identified the key drivers of cardiovascular risk in the dataset. Age stands out as the dominant predictor, followed by systolic blood pressure, smoking status, and gender, with body composition metrics (weight, BMI, height²) playing supporting roles. The model provides a robust framework for understanding CVD risk factors and could be applied to risk stratification in clinical or public health settings. All predictors demonstrate statistical significance and contribute meaningfully to the model's predictive capability, which confirms that the backward elimination process yielded an optimal balance between model complexity and explanatory power.