A logo for university computing

Auto-generated description

**Evaluation cover page**

|  |  |
| --- | --- |
| *Student's Full Name:* | Taylan Ozgur Ozkan |
| *Student Number:* | 2024140 |
| *Module Title:* | Statistical Techniques for Data Analysis |
| *Evaluation Title:* | CA1 |
| *Assessment Due Date:* | 21/04/2024. |
| *Presentation Date* | 26/04/2024. |

**Declaration**

By submitting this review, I confirm that I have read CCT's policy on academic misconduct and understand the implications of submitting work that is not mine or that does not appropriately reference material taken from a third party or other source.

I declare that it is my own work and that all third-party material has been properly referenced.

I further confirm that this work has not previously been submitted for evaluation by me or anyone else at CCT College Dublin or any other higher education institution.

**Contents**

[**Review and Clean Dataset:** 1](#_Toc165059201)

[**1. Uploading the Data Set and Initial Review:** 1](#_Toc165059202)

[**2. Examining and Filling in Missing Values:** 1](#_Toc165059203)

[**3. Conversion of Data Types:** 1](#_Toc165059204)

[**4. Data Set After Changes:** 1](#_Toc165059205)

[**Descriptive Statistics for Numerical Variables:** 1](#_Toc165059206)

[**Frequency Tables for Categorical Variables:** 2](#_Toc165059207)

[**Visualization** 3](#_Toc165059208)

[**Histograms for Numerical Variables:** 3](#_Toc165059209)

[**Bar Charts for Categorical Variables:** 4](#_Toc165059210)

[**Graphs of other variable comparisons:** 5](#_Toc165059211)

[**Statistics** 6](#_Toc165059212)

[**Statistics Grouped by Gender:** 6](#_Toc165059213)

[**Statistics Grouped by Smoking Status:** 6](#_Toc165059214)

[**5. Results and Discussion** 7](#_Toc165059215)

The Impact of Smoking on Health

# **Review and Clean Dataset:**

## **1. Uploading the Data Set and Initial Review:**

• The data set consists of 3900 rows and 7 columns.

• Columns: age, sex, current\_smoker, heart\_rate, blood\_pressure, cigs\_per\_day, chol.

• I showed the first five lines.

## **2. Examining and Filling in Missing Values:**

• 14 and 7 missing values were found in cigs\_per\_day and chol columns, respectively.

• I filled in these missing values with the averages of these columns respectively.

## **3. Conversion of Data Types:**

• I created two new numerical columns as systolic\_bp and diastolic\_bp by dividing the blood\_pressure column into two according to the '/' sign and converted these values to float type.

## **4. Data Set After Changes:**

• I have shown the first five rows, along with new columns and updated missing values.

# **Descriptive Statistics for Numerical Variables:**

• **Age (age):** Mean ~49.5, Standard Deviation ~8.56, Minimum 32, Maximum 70

• **Heart Rate (heart\_rate):** Average ~75.7, Standard Deviation ~12.02, Minimum 44, Maximum 143

**• Number of Cigarettes Consumed per Day (cigs\_per\_day):** Average ~9.17, Standard Deviation ~12.02, Minimum 0, Maximum 70

**• Cholesterol Level (chol):** Mean ~236.6, Standard Deviation ~44.34, Minimum 113, Maximum 696

**• Systolic Blood Pressure (systolic\_bp):** Mean ~132.4, Standard Deviation ~21.98, Minimum 83.5, Maximum 295

**• Diastolic Blood Pressure (diastolic\_bp):** Mean ~83, Standard Deviation ~11.92, Minimum 48, Maximum 142.5

# **Frequency Tables for Categorical Variables:**

**• Gender (sex):**

• Female: 2081

• Male: 1819

**• Smoking (current\_smoker):**

• Non-smoker (no): 1968

• Smoker (yes): 1932

# **Visualization**

## **Histograms for Numerical Variables:**

**A group of graphs showing different colors

Description automatically generated**

• **Age**: There is generally a concentration in the 40-60 age range.

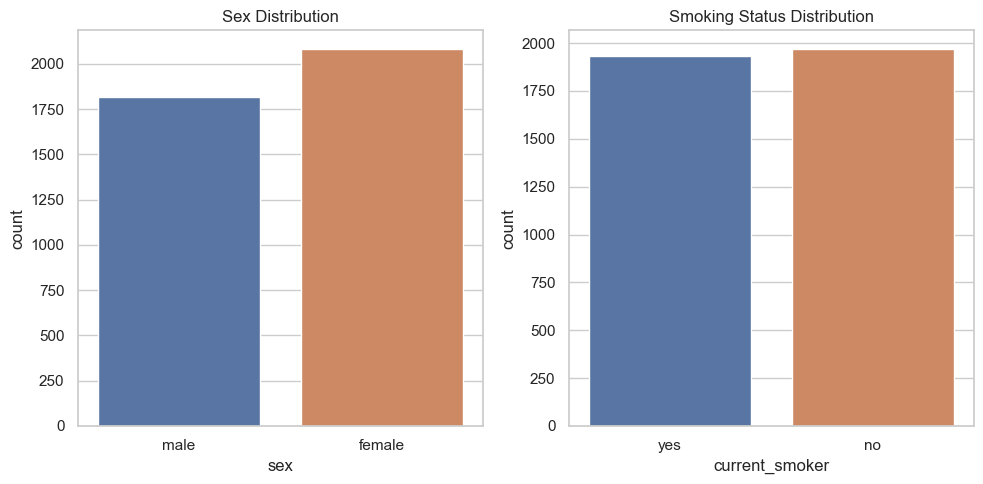
• **Heart Rate**: The distribution is mostly concentrated between 60-90 bpm.

• **Number of Cigarettes Consumed Per Day**: While most individuals do not smoke at all, some individuals consume high amounts of cigarettes.

• **Cholesterol Level**: It varies between 200-300 mg/dL.

Systolic Blood Pressure and Diastolic Blood Pressure: Both measurements are distributed within ranges considered normal for healthy adults.

## **Bar Charts for Categorical Variables:**



• Gender (Sex): There are slightly more women than men in the data set.

• Smoking (Current Smoker): The numbers of smokers and non-smokers are almost equal.

## **Graphs of other variable comparisons:**

A graph of different colored bars

Description automatically generated with medium confidence

A graph of different sizes and colors

Description automatically generated with medium confidence

A graph of a number of people

Description automatically generated with medium confidence

A graph of numbers and a number of people

Description automatically generated with medium confidence

# **Statistics**

## **Statistics Grouped by Gender:**

**• Women (Female):**

• Average Age: ~49.76

• Average Heart Rate: ~76.94 bpm

• Average Daily Cigarette Consumption: ~5.53

• Average Cholesterol Level: ~239.63 mg/dL

• Mean Systolic Blood Pressure: ~133.25 mmHg

• Average Diastolic Blood Pressure: ~82.36 mmHg

**• Men (Male):**

• Average Age: ~49.29

• Average Heart Rate: ~74.26 bpm

• Average Daily Cigarette Consumption: ~13.33

• Average Cholesterol Level: ~233.13 mg/dL

• Mean Systolic Blood Pressure: ~131.42 mmHg

• Average Diastolic Blood Pressure: ~83.70 mmHg

## **Statistics Grouped by Smoking Status:**

**• Non-Smokers (No):**

• Average Age: ~51.38

• Average Heart Rate: ~75.01 bpm

• Average Daily Cigarette Consumption: 0 (Non-smokers)

• Average Cholesterol Level: ~238.64 mg/dL

• Average Systolic Blood Pressure: ~135.34 mmHg

• Average Diastolic Blood Pressure: ~84.31 mmHg

**• Smokers (Yes):**

• Average Age: ~47.66

• Average Heart Rate: ~76.38 bpm

• Average Daily Cigarette Consumption: ~18.51

• Average Cholesterol Level: ~234.51 mg/dL

• Average Systolic Blood Pressure: ~129.39 mmHg

• Average Diastolic Blood Pressure: ~81.64 mmHg

These results show differences and relationships between various groups in the data set.

# **5. Results and Discussion**

**Evaluation of Findings**: Our analyzes showed that smoking has some effect on heart rate. Heart rate was higher in smokers than in non-smokers. Additionally, a positive correlation was found between age and cholesterol level and heart rate.

**Limitations and Recommendations**: The limitations of this study include the insufficient diversity of the data set used and the limited number of variables used in the analysis. It is recommended to include more variables in future studies and conduct similar analyzes on different population groups.