KING'S OWN INSTITUTE*

Success in Higher Education

INDIVIDUAL ASSIGNMENT COVERSHEET

Family Name: TASNIM Given Name: SABABA

Student Number: 20028086 Lecturer's/ Tutor`s Name: Firoz ANWAR

Subject Code & Name: ICT370 Data Analytics T225

Assignment Title: Individual Progress Report and Reflection-Submission Week 6

Declaration

(This declaration must be completed by the student or the assignment will not be marked.)

I certify the following:

- I have read and understood the *Student Academic Misconduct Policy*.
- This assignment is my own work based on my personal study and or research.
- I have acknowledged all material and sources used in the preparation of this assignment including any material generated in the course of my employment.
- The assignment has not previously been submitted for assessment.
- I have not copied in part or in whole or otherwise plagiarised the work of other students.
- I have read and I understand the criteria used for assessment.
- The assignment is within the word and page limits specified in the unit outline.
- The use of any material in this assignment does not infringe the intellectual property / copyright of a third party.
- I understand that this assignment may undergo electronic detection for plagiarism, and an anonymous copy of the assignment may be retained on the database and used to make comparisons with other assignments in future.
- By completing this coversheet in full and submitting this assignment electronically, I am bound by the conditions of the KOI's *Student Academic Misconduct Policy* and the declaration on this coversheet.

sababa
Signature

07 / 08 / 2025
Date

Assignment Receipt

Family Name: TASNIM Given Name: SABABA

Student Number: 20028086 Lecturer's/ Tutor`s Name: Firoz ANWAR

Subject Code & Name: ICT370 Data Analytics

Assignment Title: Individual Progress Report and Reflection-Submission Week 6

sababa
Signature

07 / 08 / 2025
Date

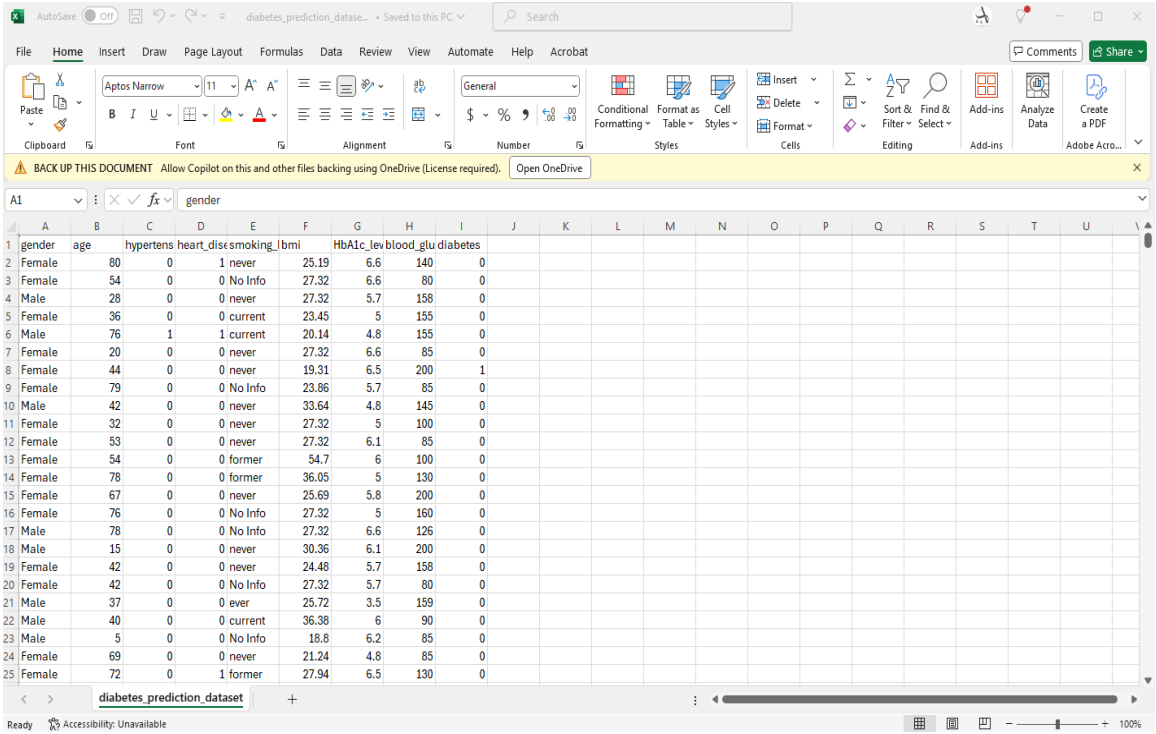
Introduction:

In this report, we focus on the diabetes-prediction-dataset (100,000 rows × 9 columns) dataset. As part of the analysis, we compute the measures of central tendency and dispersion as well as the distribution of all the numeric columns with histograms and boxplots. We also analyze anomalies, (missing data, outliers) and provide recommendations. Capture evidence of work, justified methods, and document progress for every week worked on the project.

Central Tendency and Dispersion:

The following table summarizes mean, median, std dev, skewness, and outlier % for each numeric column:

Variable	mean	median	stddev_s	skewness	outliers_pct
age	41.886	43.0	22.517	-0.052	0.0
hypertension	0.075	0.0	0.263	3.231	7.485
heart_disease	0.039	0.0	0.195	4.734	3.942
bmi	27.321	27.32	6.637	1.044	7.086
HbA1c_level	5.528	5.8	1.071	-0.067	1.315
blood_glucose_level	138.058	140.0	40.708	0.822	2.038
diabetes	0.085	0.0	0.279	2.976	8.5



gender	age	hypertens	heart_dise	smoking_l	bmi	HbA1c_lev	blood_glu	diabetes
Female	80	0	1	never	25.19	6.6	140	0
Female	54	0	0	No Info	27.32	6.6	80	0
Male	28	0	0	never	27.32	5.7	158	0
Female	36	0	0	current	23.45	5	155	0
Male	76	1	1	current	20.14	4.8	155	0
Female	20	0	0	never	27.32	6.6	85	0
Female	44	0	0	never	19.31	6.5	200	1
Female	79	0	0	No Info	23.86	5.7	85	0
Male	42	0	0	never	33.64	4.8	145	0
Female	32	0	0	never	27.32	5	100	0
Female	53	0	0	never	27.32	6.1	85	0
Female	54	0	0	former	54.7	6	100	0
Female	78	0	0	former	36.05	5	130	0
Female	67	0	0	never	25.69	5.8	200	0
Female	76	0	0	No Info	27.32	5	160	0
Male	78	0	0	No Info	27.32	6.6	126	0
Male	15	0	0	never	30.36	6.1	200	0
Female	42	0	0	never	24.48	5.7	158	0
Female	42	0	0	No Info	27.32	5.7	80	0
Male	37	0	0	ever	25.72	3.5	159	0
Male	40	0	0	current	36.38	6	90	0
Male	5	0	0	No Info	18.8	6.2	85	0
Female	69	0	0	never	21.24	4.8	85	0
Female	72	0	1	former	27.94	6.5	130	0

Fig. 1: Excel descriptive statistics output for Age, BMI, HbA1c, Blood Glucose, etc.

Dataset Interpretations:

- **age:** For 'age', mean=41.89, median=43.00, mode=80.0.
  Here the distribution is symmetric (skew=-0.05). Spread: std≈22.52, IQR≈36.00. Outliers (Tukey) ≈ 0.00%.

- **hypertension:** For 'hypertension', mean=0.07, median=0.00, mode=0.0.
  Here the distribution is right-skewed (skew=3.23). Spread: std≈0.26, IQR≈0.00. Outliers (Tukey) ≈ 7.49%.

- **heart_disease:** For 'heart_disease', mean=0.04, median=0.00, mode=0.0.
  In this dataset, the distribution is right-skewed (skew=4.73). Spread: std≈0.19, IQR≈0.00. Outliers (Tukey) ≈ 3.94%.

- **bmi:** For 'bmi', mean=27.32, median=27.32, mode=27.32.
  Here the distribution is right-skewed (skew=1.04). Spread: std≈6.64, IQR≈5.95. Outliers (Tukey) ≈ 7.09%.

- **HbA1c_level:** For ‘HbA1c_level’, mean=5.53, median=5.80, mode=6.6. Here the distribution is approximately symmetric (skew=-0.07). Spread: std≈1.07, IQR≈1.40. Outliers (Tukey) ≈ 1.31%.

- **blood_glucose_level:** For ‘blood_glucose_level’, mean=138.06, median=140.00, mode=130.0. Here it shows the distribution is right-skewed (skew=0.82). Spread: std≈40.71, IQR≈59.00. Outliers (Tukey) ≈ 2.04%.

- **diabetes:** For ‘diabetes’, mean=0.09, median=0.00, mode=0.0. Here the distribution appears right-skewed (skew=2.98). Spread: std≈0.28, IQR≈0.00. Outliers (Tukey) ≈ 8.50%.

Histograms:

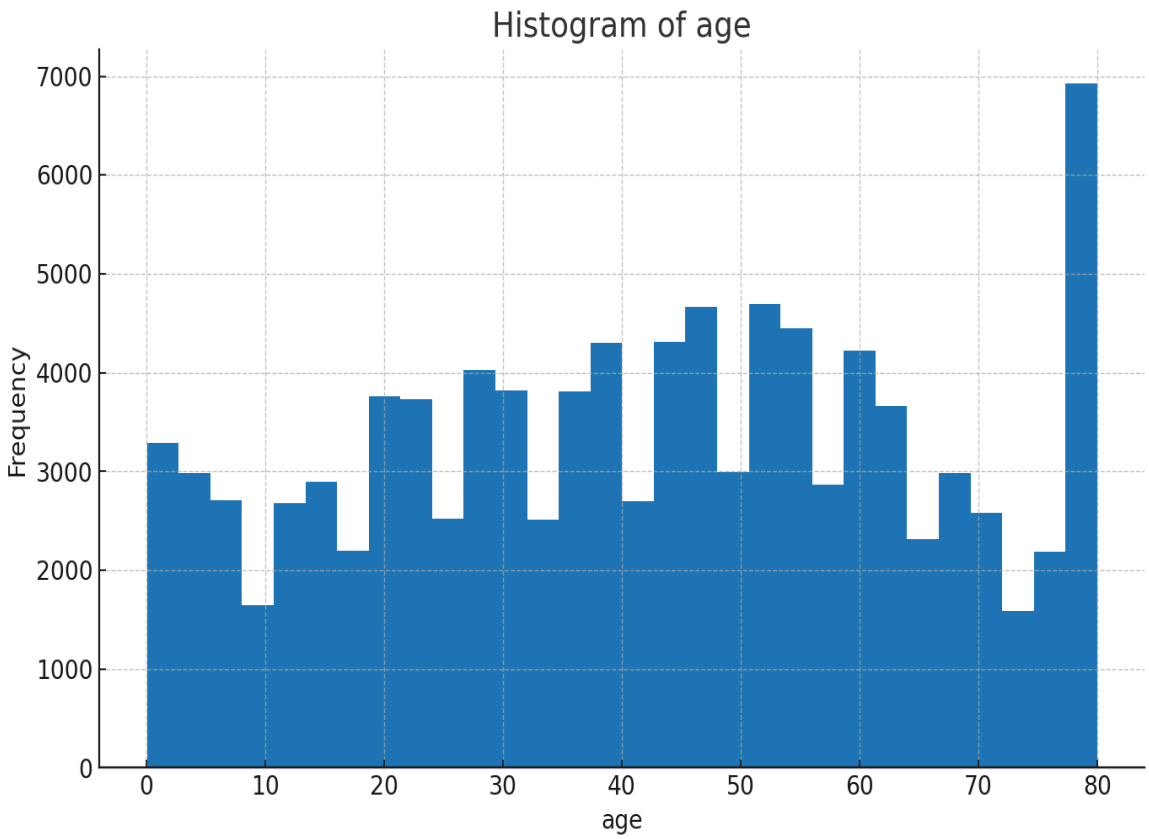


Figure 2.1: Histogram of age

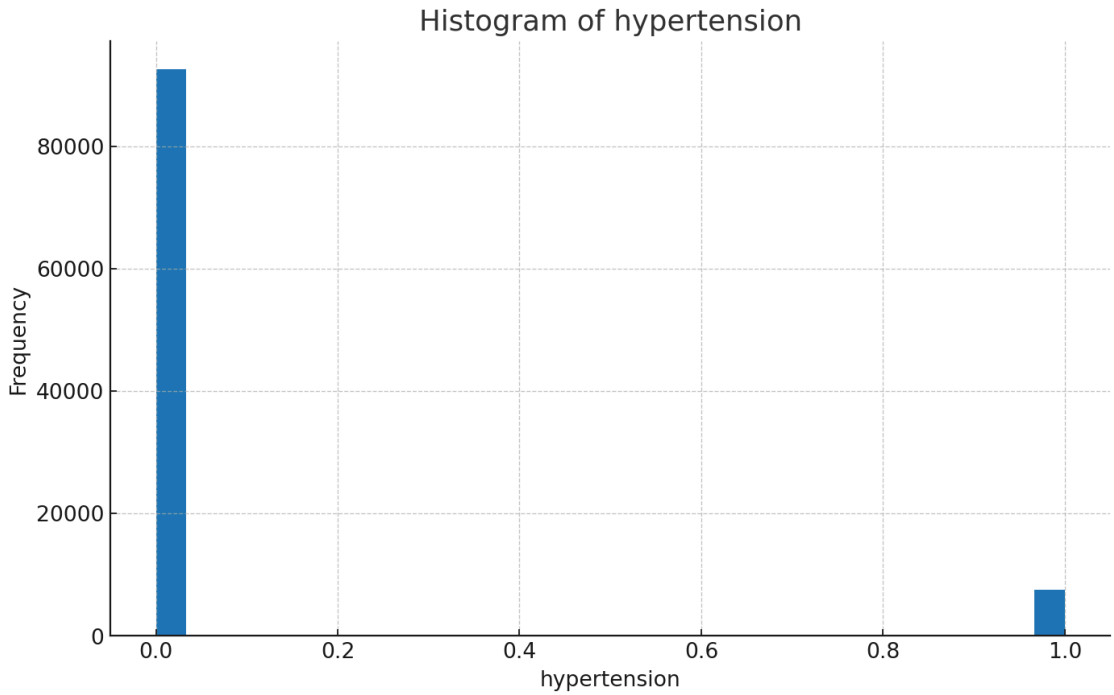


Figure 2.2: Histogram of hypertension

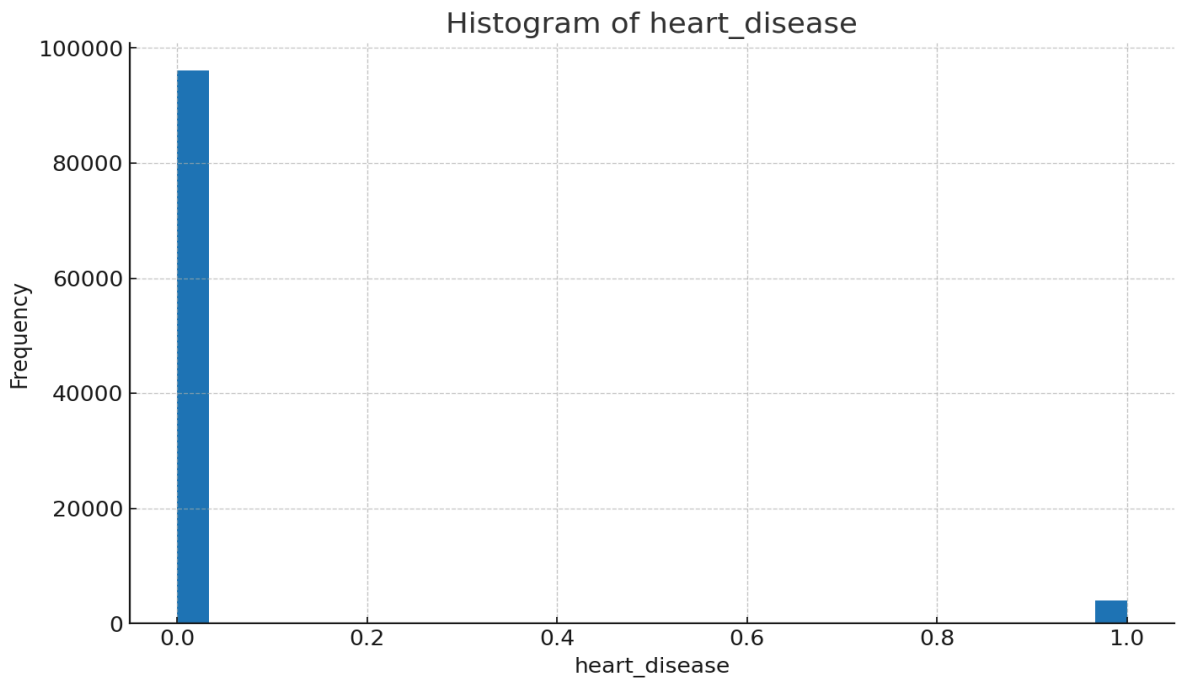


Figure 2.3: Histogram of heart_disease

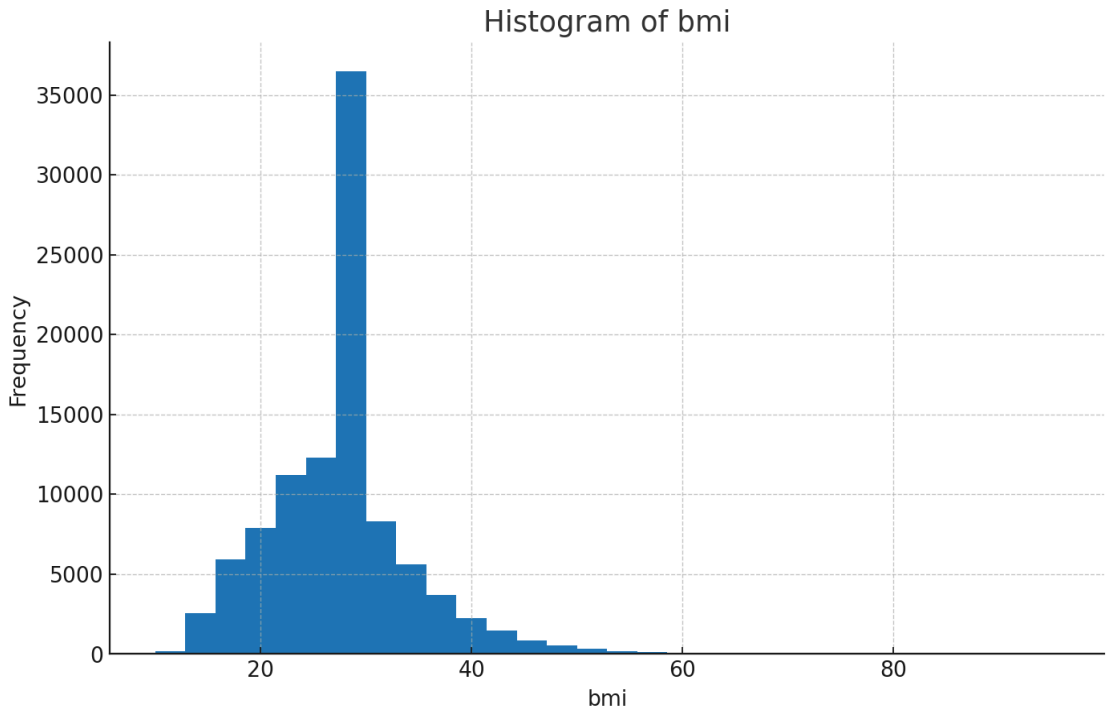


Figure 2.4: Histogram of bmi

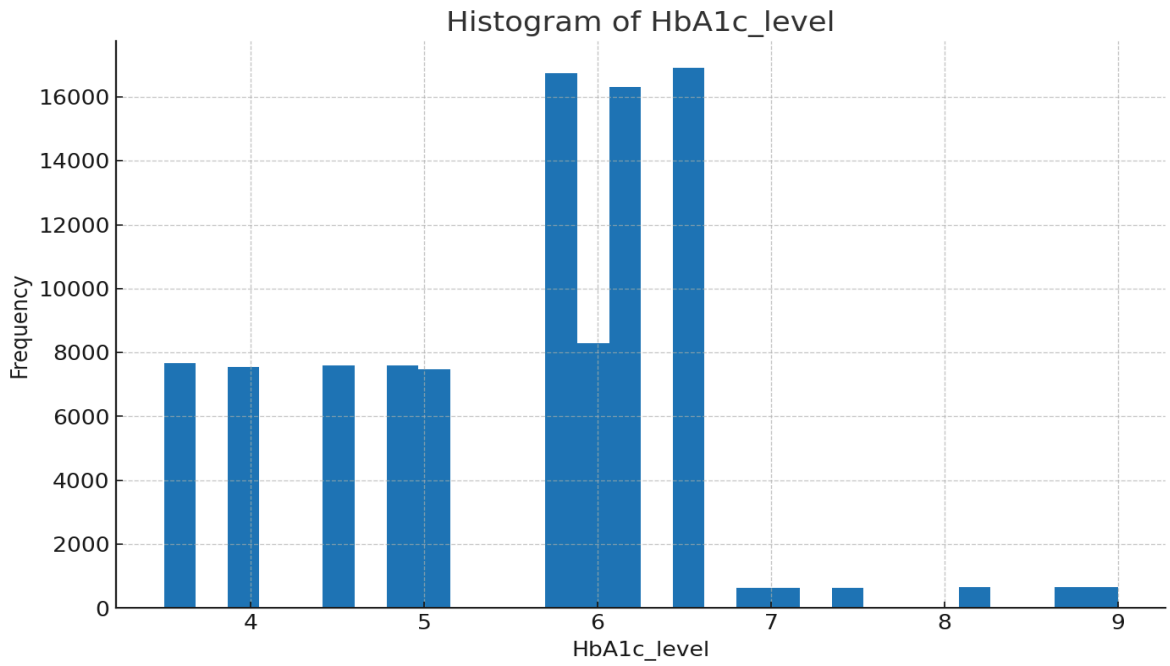


Figure 2.5: Histogram of HbA1c_level

Figure 2.6: Histogram of blood_glucose_level

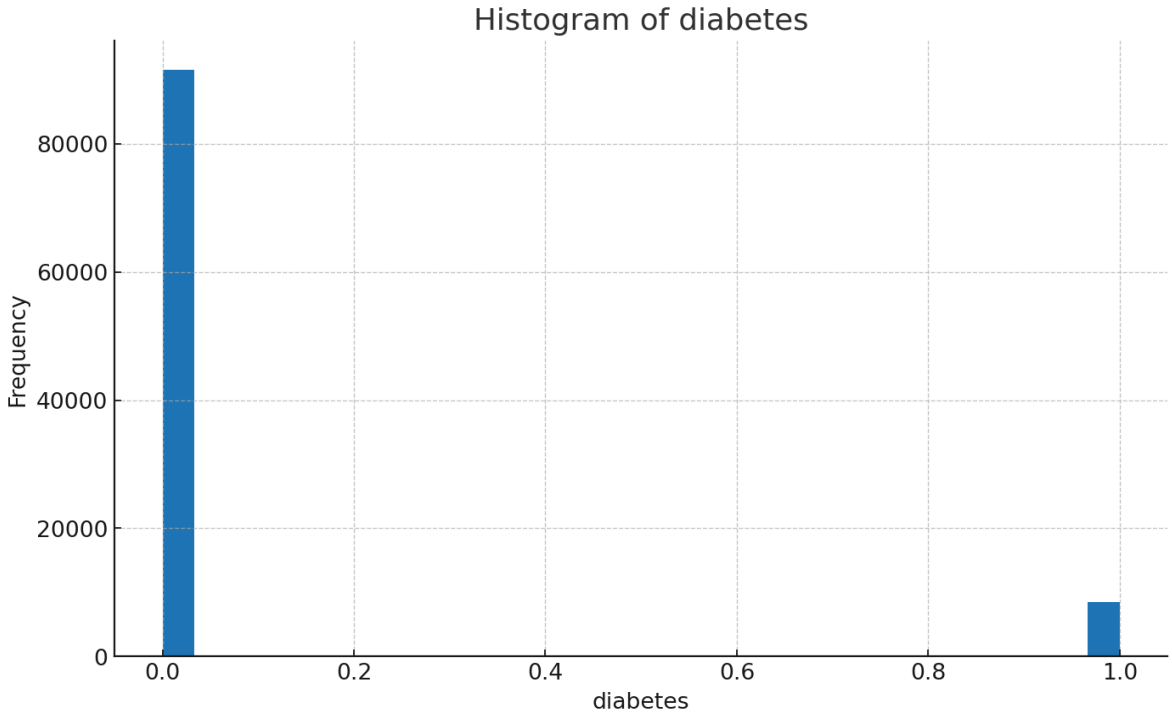


Figure 2.7: Histogram of diabetes

Box and Whisker Plots:

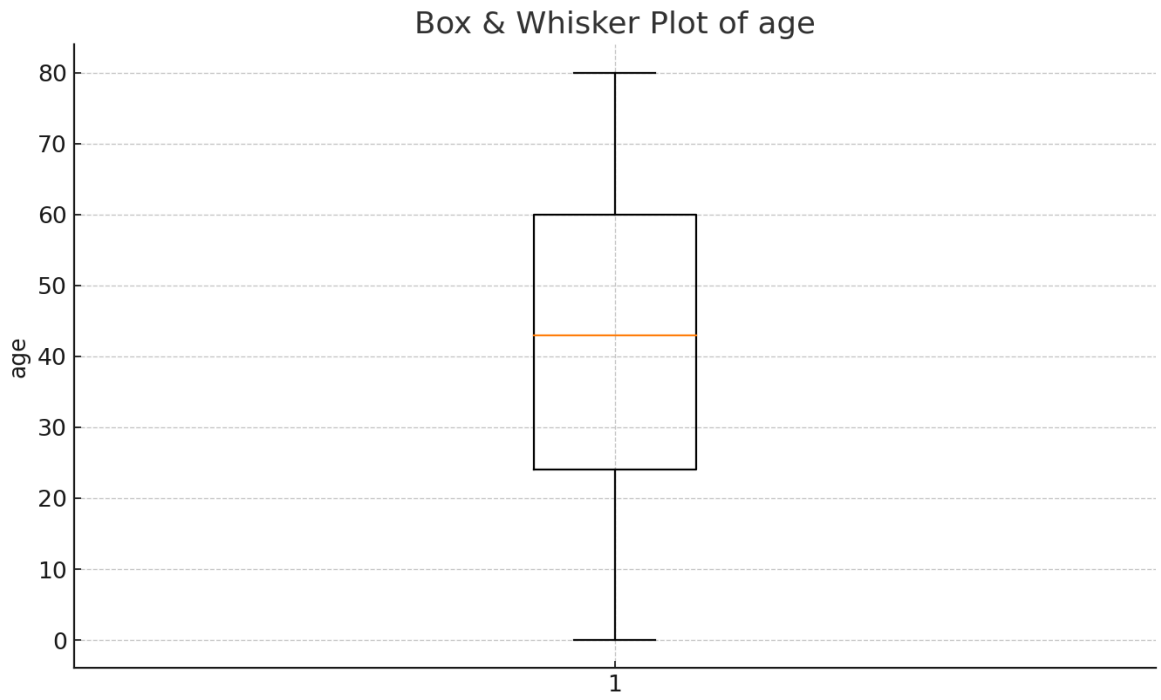


Figure 3.1: Boxplot of age

Figure 3.2: Boxplot of hypertension

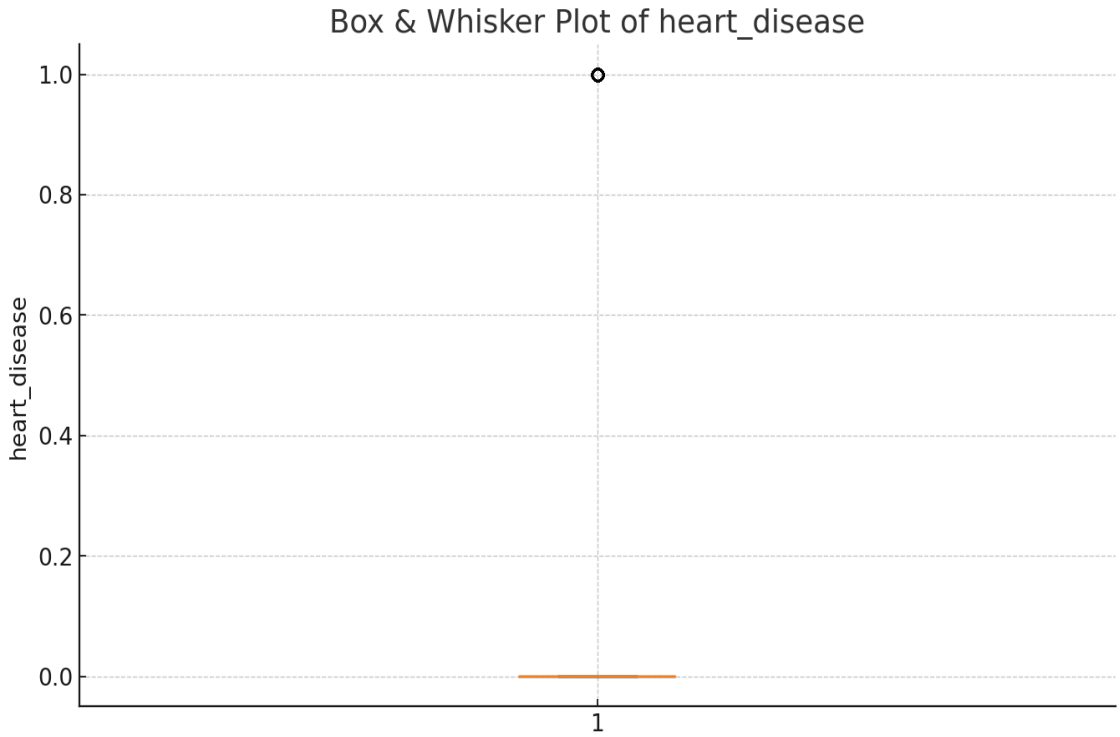


Figure 3.3: Boxplot of heart_disease

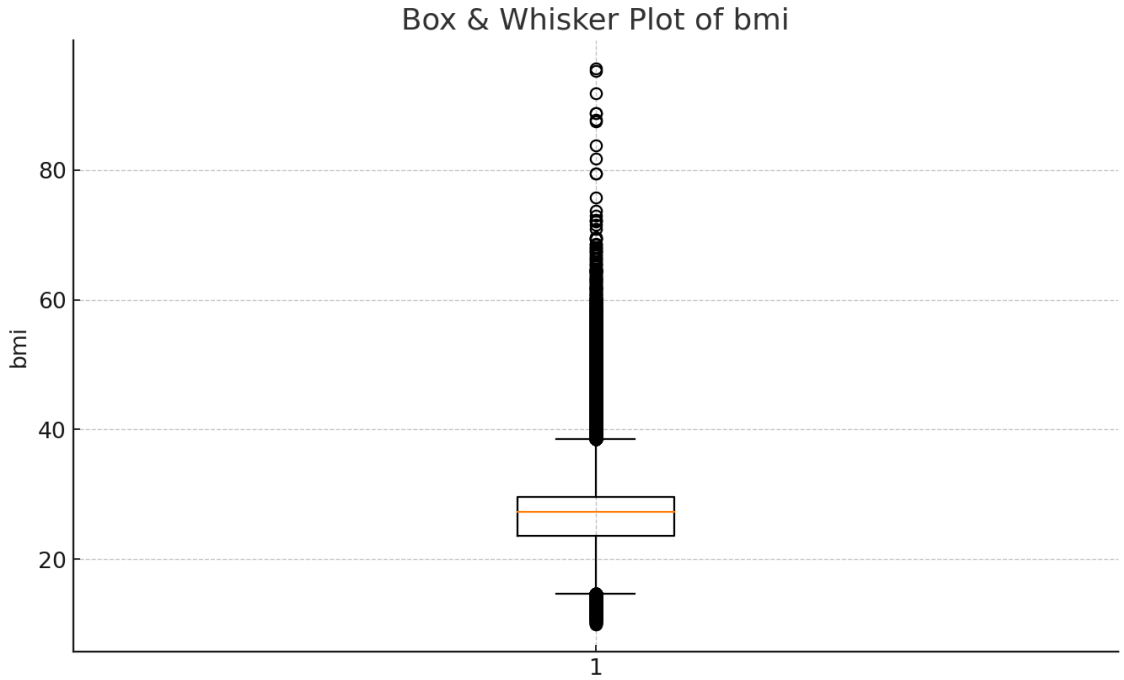


Figure 3.4: Boxplot of bmi

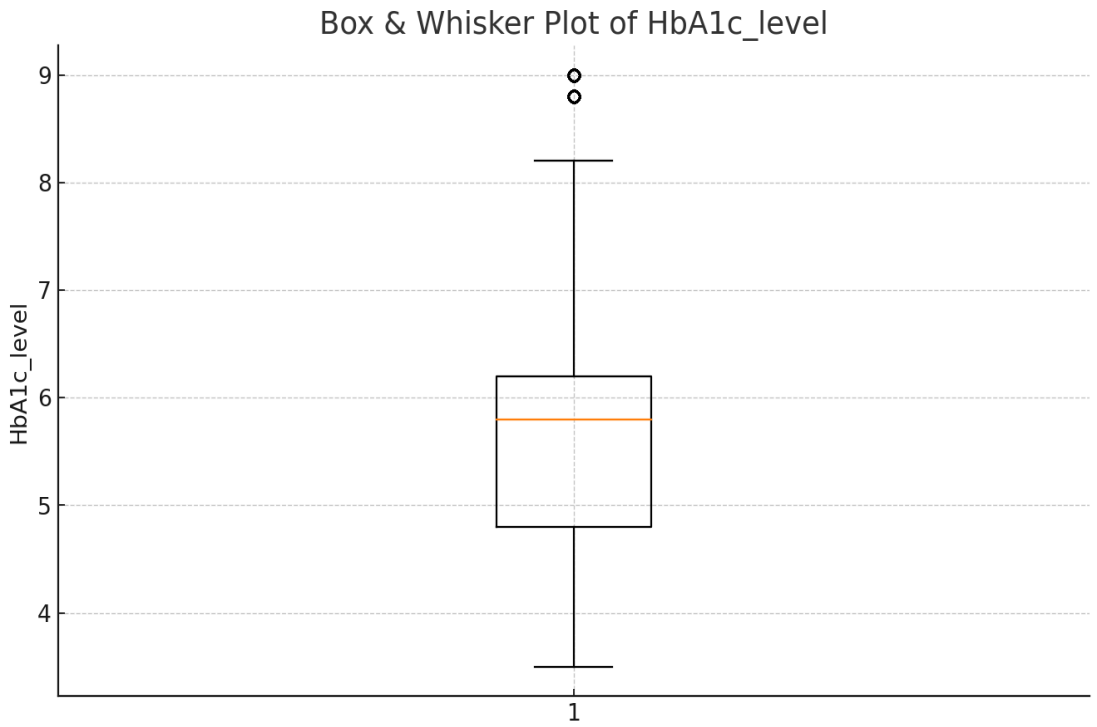


Figure 3.5: Boxplot of HbA1c_level

Figure 3.6: Boxplot of blood_glucose_level

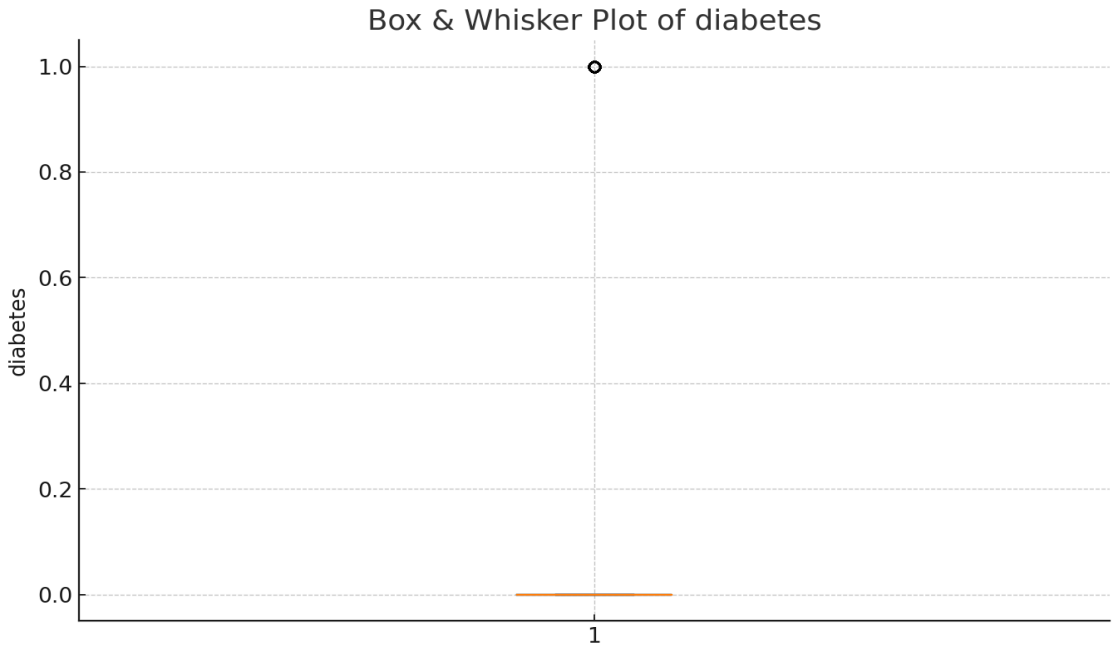


Figure 3.7: Boxplot of diabetes

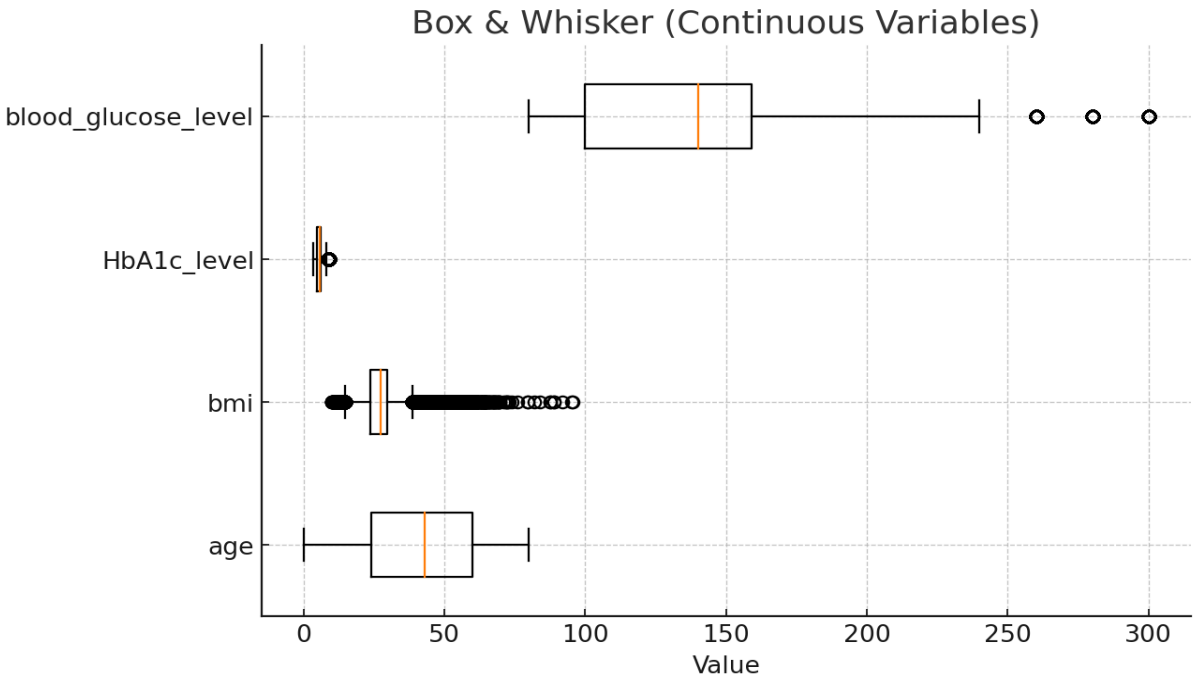
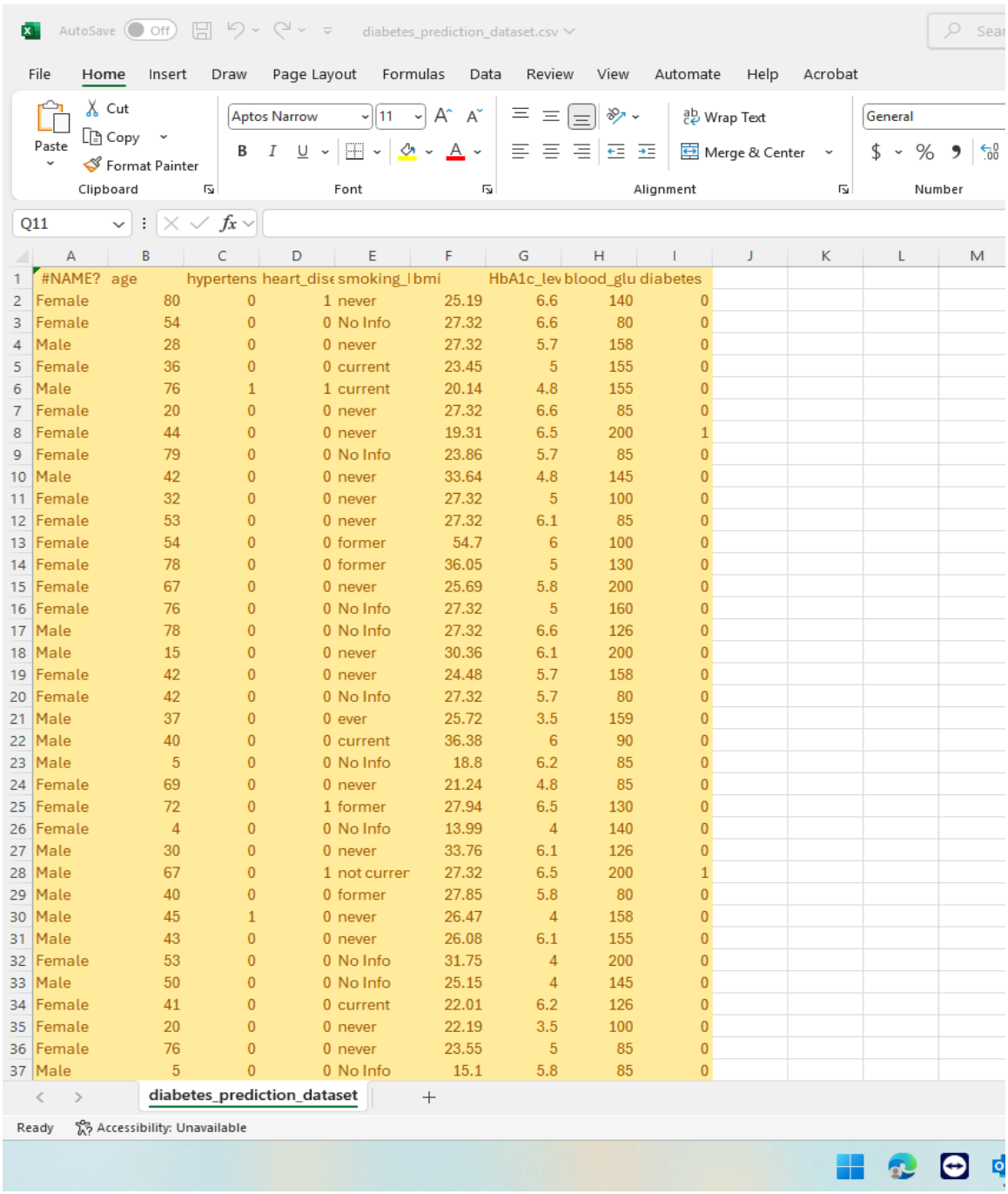


Figure 3.8: Combined Boxplots (Continuous Variables)

Anomalies & Data Quality:

There are no missing values in the dataset. Tukey's 1.5 times IQR rule was applied to identify potential outliers, where a small proportion of the continuous variables (bmi, HbA1clevel, blood-glucose-level) were flagged as outliers. However, for the binary variables (hypertension, heart-disease, diabetes), their IQR is not meaningful and should not be interpreted.



#NAME?	age	hypertens	heart_dise	smoking_l	bmi	HbA1c_lev	blood_glu	diabetes
Female	80	0	1	never	25.19	6.6	140	0
Female	54	0	0	No Info	27.32	6.6	80	0
Male	28	0	0	never	27.32	5.7	158	0
Female	36	0	0	current	23.45	5	155	0
Male	76	1	1	current	20.14	4.8	155	0
Female	20	0	0	never	27.32	6.6	85	0
Female	44	0	0	never	19.31	6.5	200	1
Female	79	0	0	No Info	23.86	5.7	85	0
Male	42	0	0	never	33.64	4.8	145	0
Female	32	0	0	never	27.32	5	100	0
Female	53	0	0	never	27.32	6.1	85	0
Female	54	0	0	former	54.7	6	100	0
Female	78	0	0	former	36.05	5	130	0
Female	67	0	0	never	25.69	5.8	200	0
Female	76	0	0	No Info	27.32	5	160	0
Male	78	0	0	No Info	27.32	6.6	126	0
Male	15	0	0	never	30.36	6.1	200	0
Female	42	0	0	never	24.48	5.7	158	0
Female	42	0	0	No Info	27.32	5.7	80	0
Male	37	0	0	ever	25.72	3.5	159	0
Male	40	0	0	current	36.38	6	90	0
Male	5	0	0	No Info	18.8	6.2	85	0
Female	69	0	0	never	21.24	4.8	85	0
Female	72	0	1	former	27.94	6.5	130	0
Female	4	0	0	No Info	13.99	4	140	0
Male	30	0	0	never	33.76	6.1	126	0
Male	67	0	1	not currer	27.32	6.5	200	1
Male	40	0	0	former	27.85	5.8	80	0
Male	45	1	0	never	26.47	4	158	0
Male	43	0	0	never	26.08	6.1	155	0
Female	53	0	0	No Info	31.75	4	200	0
Male	50	0	0	No Info	25.15	4	145	0
Female	41	0	0	current	22.01	6.2	126	0
Female	20	0	0	never	22.19	3.5	100	0
Female	76	0	0	never	23.55	5	85	0
Male	5	0	0	No Info	15.1	5.8	85	0

Figure 4: Screenshot of dataset

Recommendations:

• No actions are required regarding missing data

• Assess extreme values of BMI and glucose levels; for extreme values, consider retention, removal, or winsorization based on their plausibility.

• Assess for skew and consider log transforms.

• Apply median and IQR as opposed to mean and standard deviation when skew, outliers, or both are present.

Week-by-Week Progress

Week 1:	The focus was on data ingestion and audit. Missingness, range, and data types, as well as performing some preliminary exploratory data analysis.
Week 2:	Completing the rest of the report with all the other outstanding tasks such as boxplots, describing anomalies, other descriptive statistics, histograms, and providing recommendations.

Conclusion:

The distribution of the numeric features has different values for spread and skewness. The clinical measures of the patients like bmi, HbA1c_level, and glucose have moderate and mild skew alongside some Tukey outliers. Prevalence depicts binary indicators. There also appears to be no missing data. Based on the modeling objectives, it's advised to take into consideration the robust summaries, domain values review for extreme outliers, and sheer value of the skewed ones.

References

- Ozaydin, B., Zengul, F., Oner, N. & Feldman, S.S. 2020, "Healthcare Research and Analytics Data Infrastructure Solution: A Data Warehouse for Health Services Research", Journal of medical Internet research, vol. 22, no. 6, pp. e18579-e18579.
- Szukits, Á. (2022). The illusion of data-driven decision making–The mediating effect of digital orientation and controllers' added value in explaining organizational implications of advanced analytics. Journal of Management Control, 33(3), 403-446.
- Microsoft Excel help: AVERAGE, MEDIAN, MODE.SNGL, VAR.S, STDEV.S, QUARTILE.INC. Available here: https://support.microsoft.com/excel (Accessed: 16th August 2025).