

L21_G5

540969191 dhua0155

530419471 jzan0996

520325186 jmia6545

Auto Price

Step 1. Problem Definition

Research question

What factors influence the price of used cars, and can we build a model to predict the price (Price) from vehicle attribute.

The objective of this study is to identify the attributes that exert the greatest influence on car prices, evaluate the effectiveness of different attribute combinations in predicting price levels, and ultimately construct a predictive model that can enhance pricing accuracy and support data-driven decision-making.

Stakeholders

The results of this study are relevant to a range of stakeholders. The primary stakeholders include buyers, who can use price-influencing factors to guide negotiations and avoid mispricing; private sellers, who can highlight the most valuable selling points of their vehicles based on scientific pricing; and dealers or auctions, who can identify which configurations yield a premium and by how much they can increase prices for fully equipped cars. The secondary stakeholders include insurance companies, who can determine premiums more accurately by aligning them with vehicle risk profiles; rental car companies, who can optimize retirement and disposal strategies; and regulatory authorities, who can monitor price rationality, ensure transparency in information disclosure, and reduce information asymmetry.

Step 2. Data Description

	non_null	missing	missing_value_rate(%)	unique_values_in_cat	dtype
Unnamed: 0	18286	0	0.0	NaN	int64
Make_Model	16092	2194	12.0	9.0	object
Body_Type	16092	2194	12.0	8.0	object
Price	14995	3291	18.0	4541.0	object
Vat	11886	6400	35.0	2.0	object
Mileage	16641	1645	9.0	7507.0	object
Type	14446	3840	21.0	5.0	object
Fuel	11155	7131	39.0	4.0	object
Gears	13166	5120	28.0	NaN	float64
Comfort_Convenience	14081	4205	23.0	5289.0	object
Entertainment_Media	8595	9691	53.0	290.0	object
Extras	2743	15543	85.0	267.0	object
Safety_Security	5121	13165	72.0	1940.0	object
Age	17006	1280	7.0	NaN	float64
Previous_Owners	9875	8411	46.0	NaN	float64
Horsepower	17006	1280	7.0	84.0	object
Inspection_New	11886	6400	35.0	NaN	float64
Paint_Type	7863	10423	57.0	3.0	object
Upholstery_Type	3109	15177	83.0	2.0	object
Gearing_Type	17006	1280	7.0	3.0	object
Displacement	13349	4937	27.0	64.0	object
Weight	16275	2011	11.0	922.0	object
Drive_Chain	12252	6034	33.0	3.0	object
Cons_Comb	12435	5851	32.0	NaN	float64

(Figure 1)

The original dataset is a CSV file contains 18,286 instances and 24 attributes with mixture of object, float, and integer. A summary table (Figure 1) was constructed to present the data types of each attribute and the distribution of missing values more intuitively, it consolidates information such as attribute type, number of non-null values, and number of missing values. The overall proportion of missing values is ~31%. Missing values appears in majority of the attributes, except the sequential identifier. The categorical variables demonstrate meaningful diversity. For instance, “Make_Model”, with 9 unique categories, suggesting 9 different vehicle types; “Body_Type”, with 8 unique categories, suggesting 8 body types; “Fuel”, with 4 unique categories, suggesting 4 types of fuel.

Step 3. Data Cleaning and Processing

Based on the preliminary examination results of the data table, the data quality issues are analyzed from four perspectives below.

Redundant Features

There are redundant fields in the data table. For example, Unnamed: 0 is just an export index column without business meaning and should be deleted. This column is removed during the cleaning stage.

Incorrect data types and invalid entries

```
Price Unit statistics: {'$': 8111, '€': 4806, '£': 2078}
Mileage Unit statistics: {'mi': np.int64(7813), 'km': np.int64(8828), 'unknown': np.int64(0)}
Horsepower Unit statistics: {'kW': 17006}
Displacement Unit statistics: {'cc': 13349}
Weight Unit statistics: {'kg': 9097, 'lbs': 4829, 'g': 2349}
```

(figure 2)

	Price
5555	\$-197271.94210085942
7060	\$-142101.19042530045
8298	\$-142201.3283743091
10634	£-111787.26
15364	\$-147720.12694777024
	Mileage
3426	-406274.90 mi
9928	-582983.2883414491 km
10275	-582612.067557342 km
13100	-548427.5511934992 km
	Horsepower
3201	-540.2810353120354 kW
5797	-529.9351443872027 kW
6732	-362.92366073306397 kW
17242	-529.9351443872027 kW
	Displacement
15742	-4826.332613436321 cc
17415	-4826.332613436321 cc
	Weight
2922	-3847330.79 g
3288	-2975.2683049258267 kg
5214	-4013.8920708045007 kg
8059	-3531.1278614125813 kg
8647	-1858.539936118876 kg
9866	-5828.01 lbs

(figure 3)

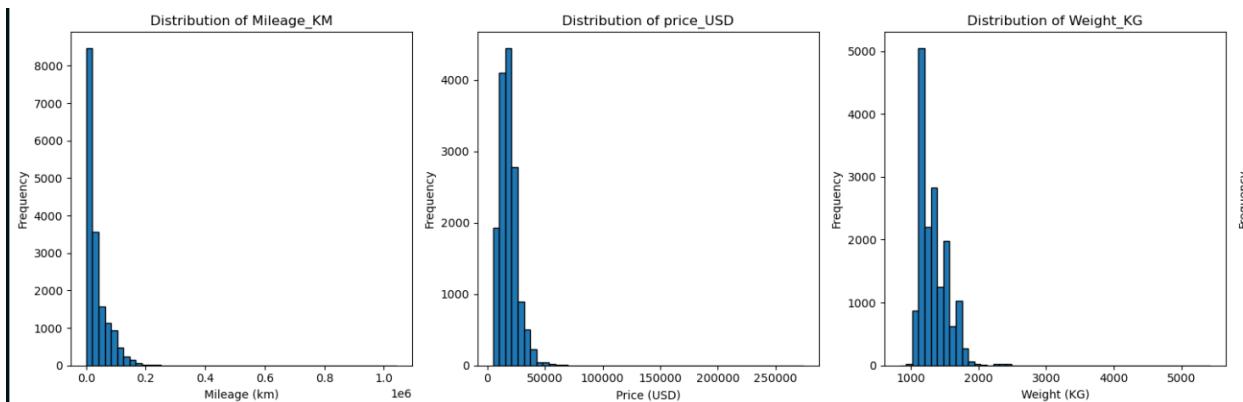
Some fields suffer from inconsistent data representation. For example, although Price (with \$/€/£), Mileage (with km/mi), Horsepower (with kW), Displacement (with cc), and Weight (with kg/lbs/g) are essentially numerical values, they are mistakenly stored as objects due to the presence of unit symbols. These are uniformly parsed into numerical values and standardized in units (see Figure 2). Several variables exhibit mismatches between their stored data types and their semantic meaning. Specifically, "Gears" and "Previous_Owners" should be encoded as integers, while "Inspection_New" should be represented as a binary

integer flag. These variables were recast into their appropriate formats to ensure consistency, and the revised attributes replaced the original fields.

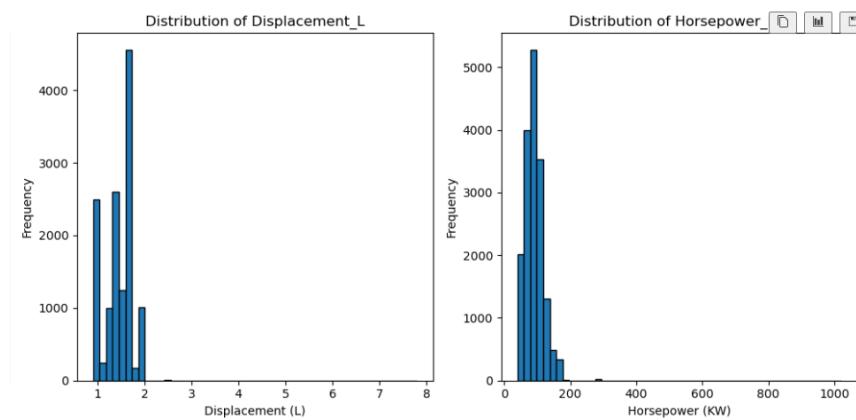
Negative values detected shall converted to NaN values as they are inconsistent with reality (see Figure 3).

Analysis of the categorical attributes revealed the presence of rare categories. To mitigate sparsity and enhance the robustness of model training, categories representing less than 0.5% of the total sample were consolidated into an “Other” category. The distributions of the categorical variables prior to consolidation are provided in the appendix(Appendix 1.1).

Missing Value Issues



(Figure 4)



(Figure 5)

A considerable proportion of missing values was detected across the dataset, with object-type attributes all exhibiting varying degrees of incompleteness. Among the 24 attributes across 18,286 records, several variables display severe missingness: "Extras" (~85%), "Safety_Security" (~72%), "Entertainment_Media" (~53%), and "Upholstery_Type" (~83%) (see Figure 1). Given their excessive missingness and limited relevance to the research objective, these variables were excluded from further analysis.

For key numerical attributes such as "Mileage_KM", "Price_USD", and "Weight_KG", missingness co-occurs with abnormal negative values. The impute of these invalid entries

are performed using the median, as these variables exhibit right-skewed distributions and the median provides robustness against extreme values (see Figures 4 and 5). For discrete numerical variables such as "Gears" and "Previous_Owners", missing values were imputed with the mode. Similarly, "Age" and "Cons_Comb" are continuous and right-skewed, and their missing values were replaced with the median. For binary variables such as "Inspection_New", mode imputation was applied to preserve class proportions.

Categorical variables were treated according to their missingness ratio. Attributes with less than 20% missing values were imputed using the mode, while those exceeding 20% were filled with an "Unknown" category. This approach balances completeness with noise reduction, while preserving the interpretability of categorical distributions.

Finally, for "Comfort_Convenience", which contains free-text descriptions of features, a new derived variable was created to record the number of listed items per observation. This transformation allows the information to be incorporated into the model in a structured and comparable manner.

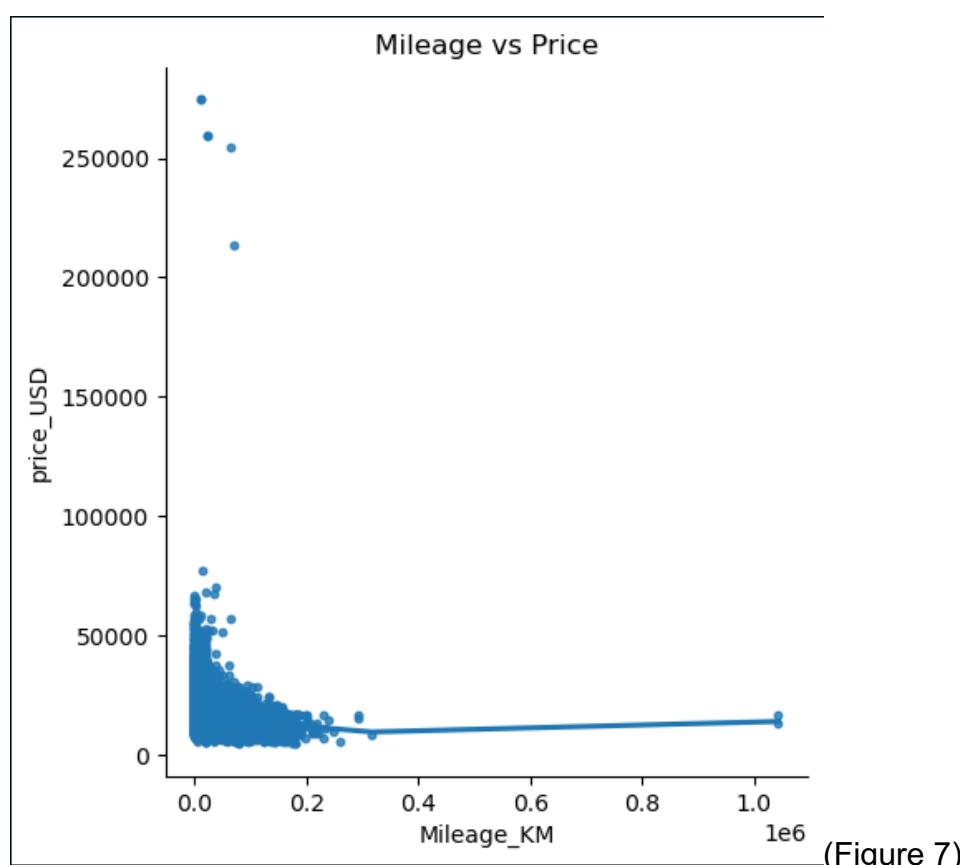
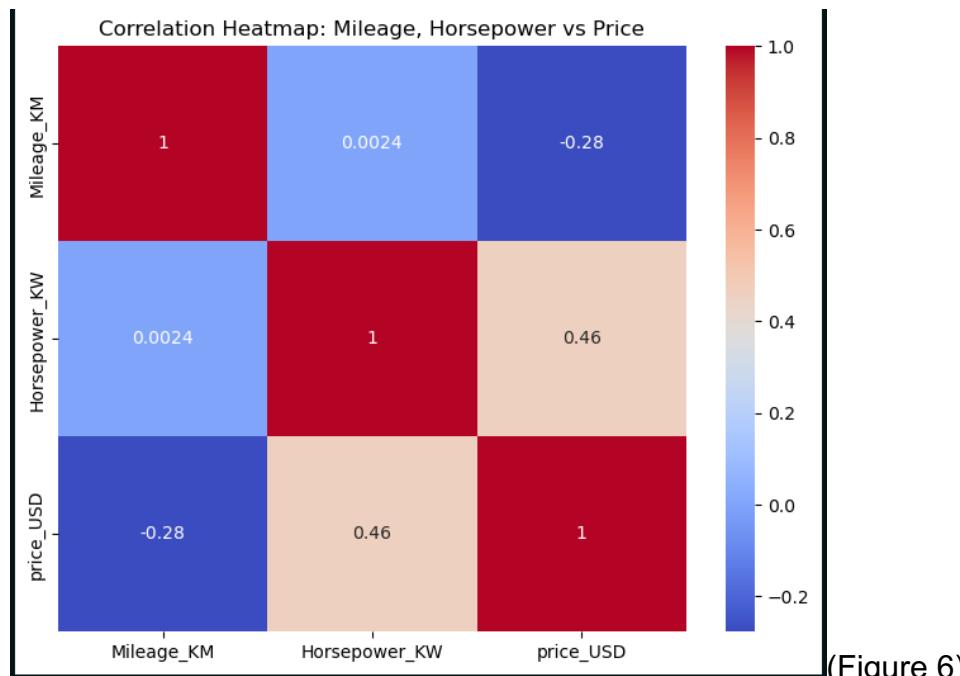
Step 4. Exploratory Data Analysis

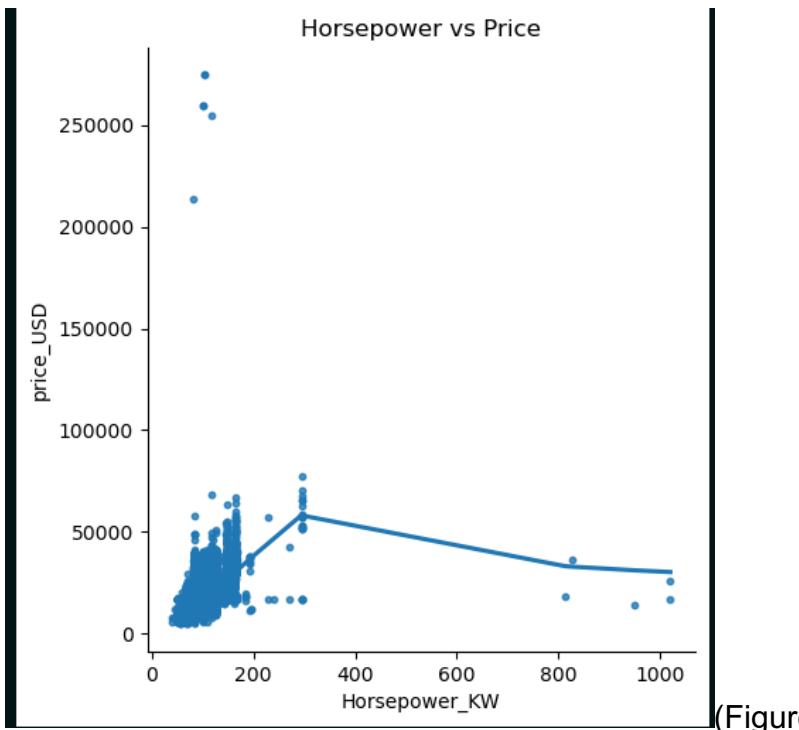
Numerical Feature Analysis

Selected Fields: Mileage_KM, Horsepower_KW

Analysis Plan: Analyze the correlation between these two fields and price using heatmaps and scatter plots.

Results:





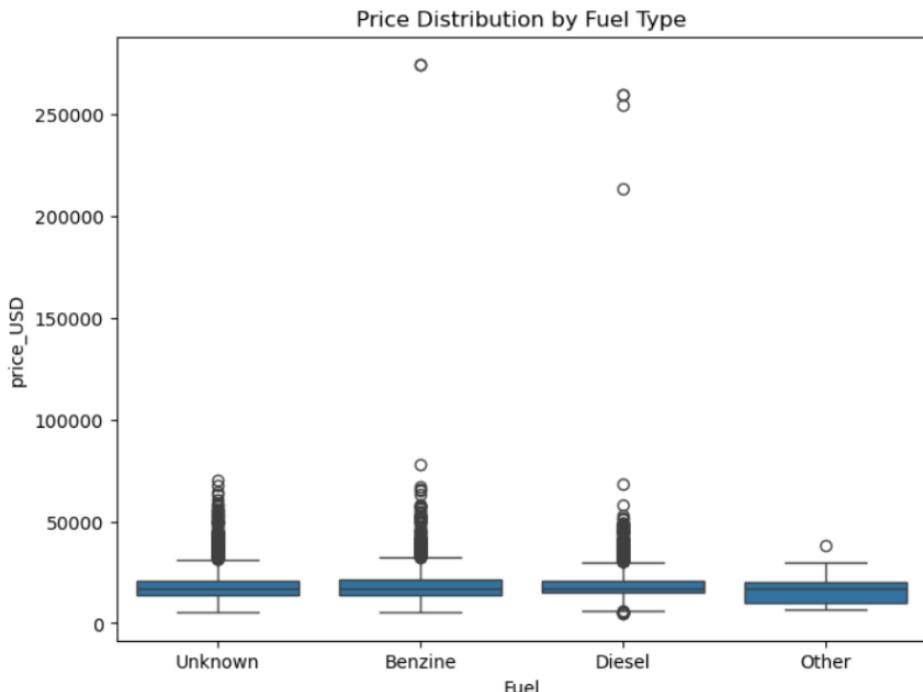
(Figure 8)

As shown in the heatmap (see Figure 6), “Mileage_KM” and “Horsepower_KW” exhibit distinct correlations with price. Notably, “Mileage_KM” shows a negative correlation with price, indicating that lower mileage corresponds to higher used car prices. Meanwhile, “Horsepower_KW” demonstrates a positive correlation with price, and the strength of this correlation is higher than that of “Mileage_KM” with “price_USD”. This suggests that vehicles with stronger power generally command higher prices.

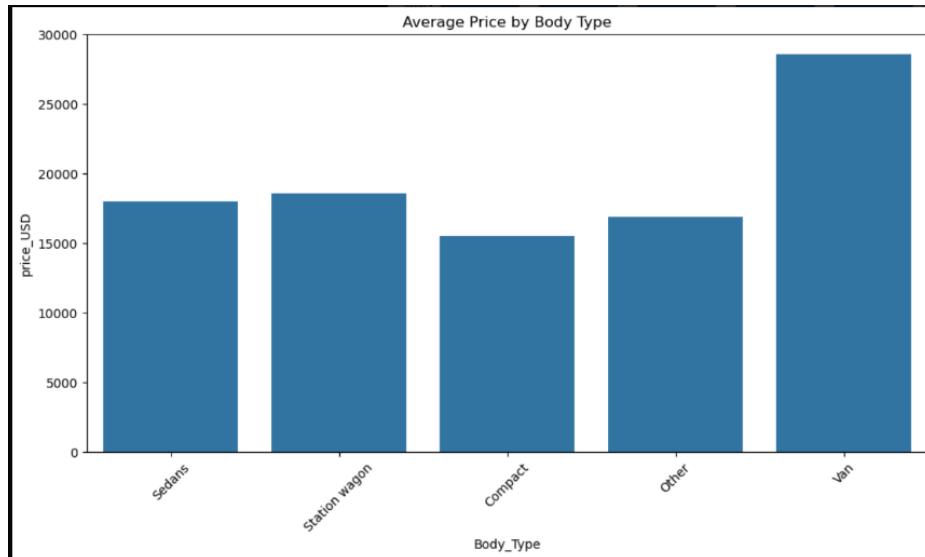
The scatter plot of “Mileage_KM” and “price_USD” (see Figure 7) shows the price decreases as mileage increases, confirming the heatmap’s conclusion: higher mileage typically leads to lower vehicle prices. When mileage is extremely high, the price stabilizes at a low level and further increases in mileage have a diminishing effect on price.

For the scatter plot of “Horsepower_KW” and “price_USD” (see Figure 8), the data roughly falls into two segments: In the lower horsepower range (approximately 0–400 KW), price increases significantly with horsepower, indicating that stronger power drives higher prices within this range. When horsepower exceeds a threshold (roughly 400 KW), price tends to decrease as horsepower increases. Here, higher horsepower no longer boosts prices, possibly due to practicality or market demand limitations. Additionally, samples with horsepower over 600 KW are sparse, indicating that extremely high-horsepower vehicles are rare in the dataset.

Categorical Feature Analysis



(Figure 9)



(Figure 10)

Selected Fields: Fuel, Body_Type

Analysis Plan: Analyze the impact of these fields on price using box plots and bar charts.

Results:

As shown in the boxplot (see Figure 9). From the position of the boxes, the median prices of Benzine, Diesel, Unknown, and Other are relatively close, all around USD 15,000–20,000, indicating that the fuel type itself is not the main factor determining vehicle price. The box sizes (interquartile ranges) of the four fuel categories show little difference, meaning the price dispersion among different fuel types is similar. In the Diesel category, some observations are far above USD 200,000 and even close to USD 280,000. This could be due to a small number of luxury diesel vehicles (e.g., high-end SUVs or commercial vehicles).

As shown in the bar chart (see Figure 9). The average prices across different body types (Body_Type) vary significantly, indicating that body design and intended use are closely related to vehicle pricing. Vans have the highest average price, nearly USD 30,000, making them the most expensive among all types, possibly due to their load capacity/space and special usage (commercial/multi-passenger vehicles). Sedans and station wagons fall in the middle, with average prices around USD 18,000–19,000, relatively stable. This suggests that these traditional body types belong to the mid-range price segment in the used car market. Compacts are the lowest, with an average price of only about USD 15,000, which aligns with expectations: small cars are positioned as economical and entry-level, depreciate quickly, and thus have lower used car prices.

Diabetes Diagnosis

Step 1. Problem Definition

Define a classification problem based on the dataset

Research Question: This study investigates whether demographic, lifestyle, and health-related factors can be used to predict an individual's diabetes status (diabetes, prediabetes, or no diabetes) in order to support prevention and intervention strategies.

Justification of Task Type: The target variable is diabetes status, with three categories: diabetes, prediabetes, and no diabetes. Since the objective is to predict membership in discrete groups rather than estimate a continuous outcome, this constitutes a classification problem.

Identification of Stakeholders and Associated Benefits

Early detection and accurate classification of diabetes status generate value for a wide range of stakeholders:

- **Patients and at-risk individuals:** Benefit from timely identification, which enables lifestyle modifications, early treatment, and a reduced likelihood of severe complications, ultimately improving quality of life.
- **Healthcare providers:** Can allocate clinical resources more effectively, prioritize high-risk groups, and design personalized care strategies.
- **Public health agencies and policymakers:** Gain insights into population-level risk factors, which inform preventive campaigns, screening initiatives, and evidence-based policy development.
- **Health insurers and payers:** Achieve cost savings by reducing long-term treatment expenditures through early management and prevention of disease progression.
- **Researchers and academic institutions:** Access valuable empirical evidence to advance scientific knowledge, refine predictive models, and design innovative interventions.

Overall, solving this classification problem supports better individual health outcomes, enhances healthcare system efficiency, reduces economic burden, and advances both policy and research agendas.

Step 2. Data Description

Number of attributes and instances

This dataset includes 253,680 examples and has 22 attributes (see Table 1). The data set's size is large enough to allow for dependable statistical analysis, and the attributes give a variety of angles to look for associations with diabetes outcomes.

Count	
Instances (rows)	264802
Attributes (columns)	23

Table 1: Summary of dataset size and attributes

Data types of each attribute

The dataset includes 23 attributes, of which 15 are numerical and 8 are categorical. Numerical attributes are mainly stored as float64, categorical attributes as object, and one column as int64. The overall distribution is summarized (Figure 1), showing that numerical variables dominate the dataset. This composition indicates that the data are well-suited for both descriptive statistics and inferential modeling. The detailed classification of each attribute is provided (Appendix 2, Table 2.1).

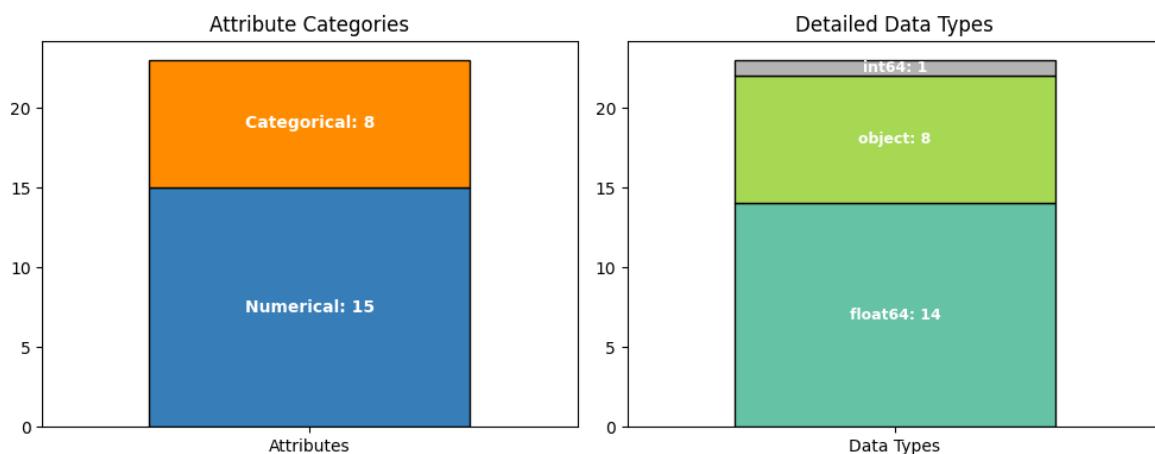


Figure 1: Distribution of Attribute Categories and Data Types

Presence and Proportion of Missing Values

Missing values are highly prevalent in this dataset, with 22 out of 23 attributes containing missing entries. The overall distribution of columns with and without missing values is shown (Figure 2). Among these, several attributes exhibit substantial missingness; the top 10 variables with the highest missing percentages are summarized (Figure 3). In particular, health-related variables such as NoDocbcCost, PhysActivity, and AnyHealthcare show more than 65% missingness, which highlights

potential challenges for downstream analysis and may require careful imputation or exclusion. A full breakdown of missing rates for all attributes is provided (Appendix 2, Table 2.2).

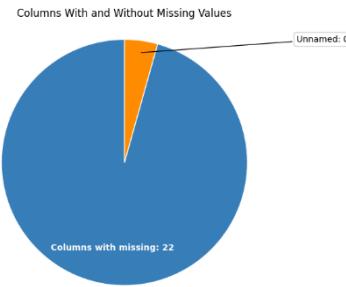


Figure 2: Attributes With and Without Missing Values

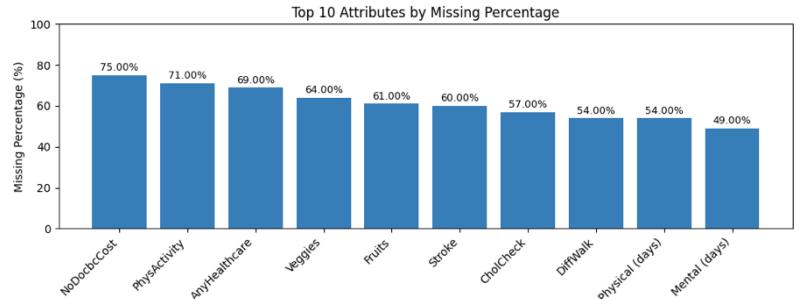


Figure 3: Top 10 Attributes by Missing Percentage

Unique Values and Distributions of Features

Categorical attributes vary in their number of unique values. For example, GeneralHealth has five levels, Education has six levels, while Diabetes contains three categories. Conversely, Income shows an unusually high number of unique values (90,184), indicating potential formatting or encoding issues. The unique values of categorical features are summarized (Table 2).

For numerical variables, distributions were assessed using horizontal boxplots, which highlight differences between binary-coded and continuous attributes. Most binary features (e.g., Smoker, CholCheck, Fruits) show highly skewed distributions concentrated at 0 and 1, while continuous variables such as BMI, Age, Physical (days), and Mental (days) exhibit wider ranges with potential outliers. The full boxplots are provided (Appendix 2, Figure 2.1).

	Attribute	Unique Values
0	GeneralHealth	5
1	Sex	2
2	Education	6
3	Income	90184
4	Diabetes	3
5	BloodPressure	2
6	Cholesterol	2
7	Alcoholic	2

Table 2: Unique Values of Categorical Attributes

Step 3. Data Cleaning and Processing

Removal of Irrelevant Columns

To ensure analytical focus, six columns were removed due to either excessive missingness (NoDocbcCost, AnyHealthcare, Physical (days), Mental (days), DiffWalk), lack of analytical relevance (Unnamed: 0), or redundancy. After this step, the dataset was reduced from 23 to 17 attributes, preserving variables most relevant to diabetes outcomes. The revised feature set includes demographic, lifestyle, and health-related factors such as BMI, Age, Education, Income, GeneralHealth, and Diabetes status, which will serve as the foundation for subsequent analysis. A summary of the column removal process is presented (Table 3).

Summary of Column Removal	
Metric	Count
Total Columns (Before)	23
Columns Removed	6
Total Columns (After)	17

Table 3: Summary of Column Removal

Formatting and Type Conversion

To prepare the dataset for consistent analysis, several formatting and type conversion steps were applied. Column names were standardized for clarity, and common placeholders (e.g., "NA", "Unknown") were replaced with NaN to unify missing values. The Income variable was converted from a string with symbols into a numeric format, enabling quantitative analysis.

Binary categorical features were recoded into 0/1 indicators, while multicategorical attributes (e.g., general_health, education, diabetes) retained their multiple levels. The diversity of categorical features is illustrated (Figure 4).

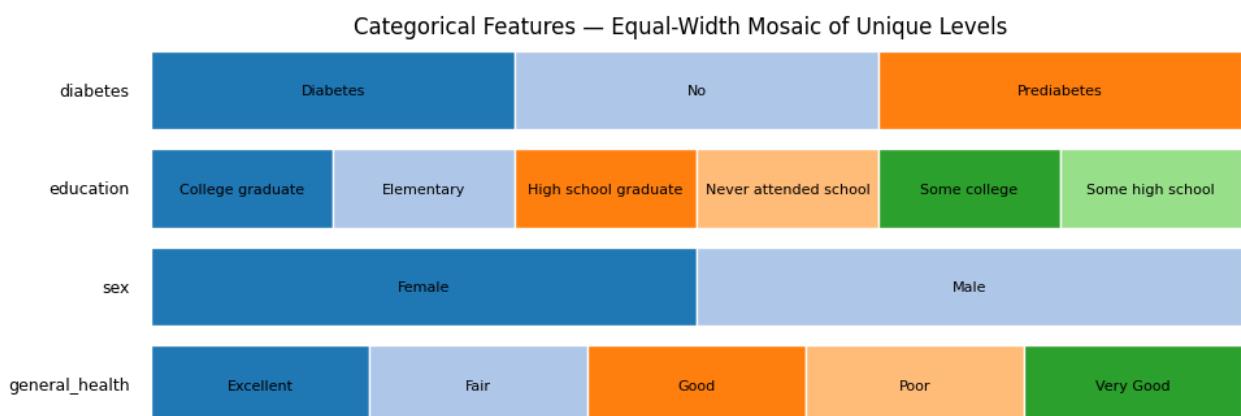


Figure 4: Unique Levels of Key Categorical Features

A comparison of attribute data types before and after formatting confirmed that several variables originally misclassified as float64 or object were corrected. Binary health indicators such as smoker, stroke, blood_pressure, and cholesterol are now stored as Int64. The complete before-and-after comparison is provided (Appendix 2, Table 2.3), while a focused summary of attributes that changed types is shown (Table 4).

	Attribute	Original Data Type	Current Data Type
0	chol_check	float64	Int64
1	smoker	float64	Int64
2	stroke	float64	Int64
3	heart_disease_attack	float64	Int64
4	phys_activity	float64	Int64
5	fruits	float64	Int64
6	veggies	float64	Int64
7	income	object	float64
8	blood_pressure	object	Int64
9	cholesterol	object	Int64
10	alcoholic	object	Int64

Table 4: Attributes with Changed Data Types Only

Handling Duplicates

The dataset was examined for exact duplicate records across all attributes. A total of 7 duplicate rows were identified and removed, ensuring that each observation represents a unique individual and preventing potential overcounting in subsequent analyses.

Detecting and Handling Outliers

To ensure data validity, potential anomalies were assessed across binary, categorical, and numerical features.

For binary categorical features, no invalid values were detected, confirming that all entries fell within the expected {0, 1, NA} range (see Appendix 2, Figure 2.2).

For non-binary categorical features, inconsistent or extraneous levels (e.g., additional "NA" codes) were identified in sex, education, general_health, and diabetes. These were mapped to <NA> to maintain consistency (Figure 5).

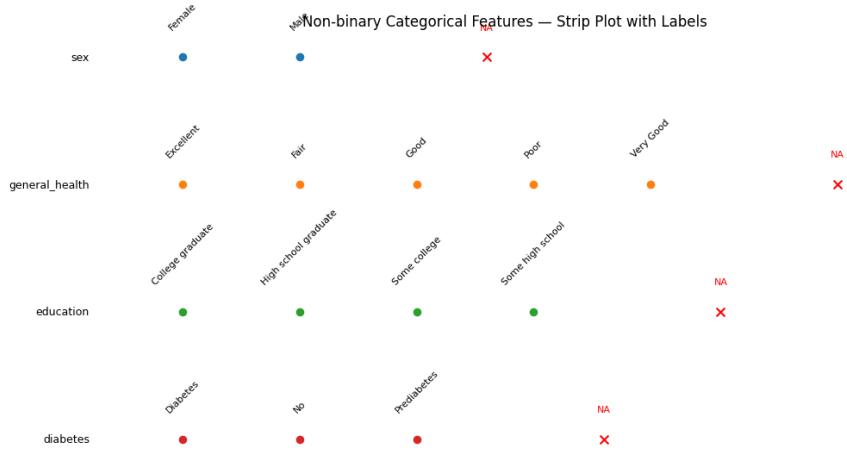


Figure 5: Distinct Levels of Non-binary Categorical Features

For numeric variables, outlier detection focused on age, BMI, and income. Their combined distribution is summarized (Figure 6), showing plausible values for age, extreme BMI cases above 50, and a highly skewed income distribution with extreme values exceeding 100,000.

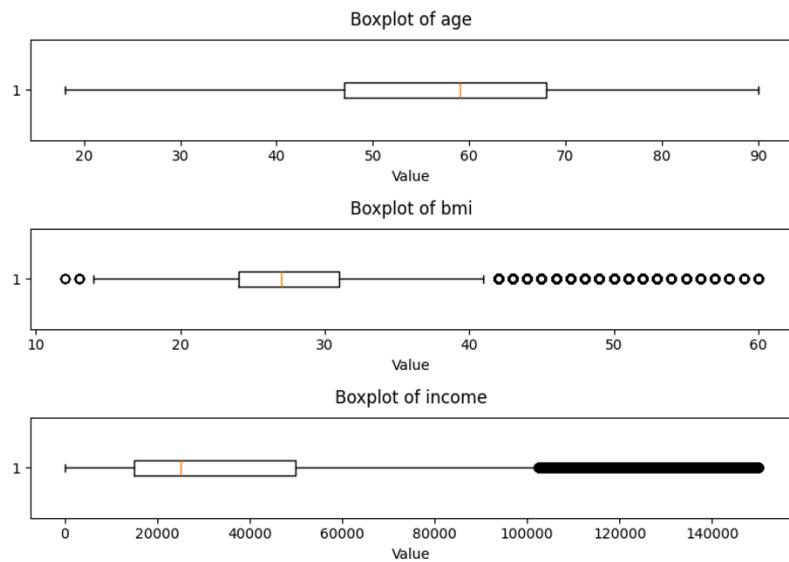


Figure 6: Boxplots of Age, BMI, and Income

Distribution of Categorical Variables

Categorical features were analyzed in two groups: binary variables and non-binary variables.

For binary categorical variables, distributions of 0, 1, and <NA> are displayed (Figure 7). The plots indicate substantial proportions of missing values for variables such as phys_activity and fruits, while others like alcoholic show strong skew toward one category.

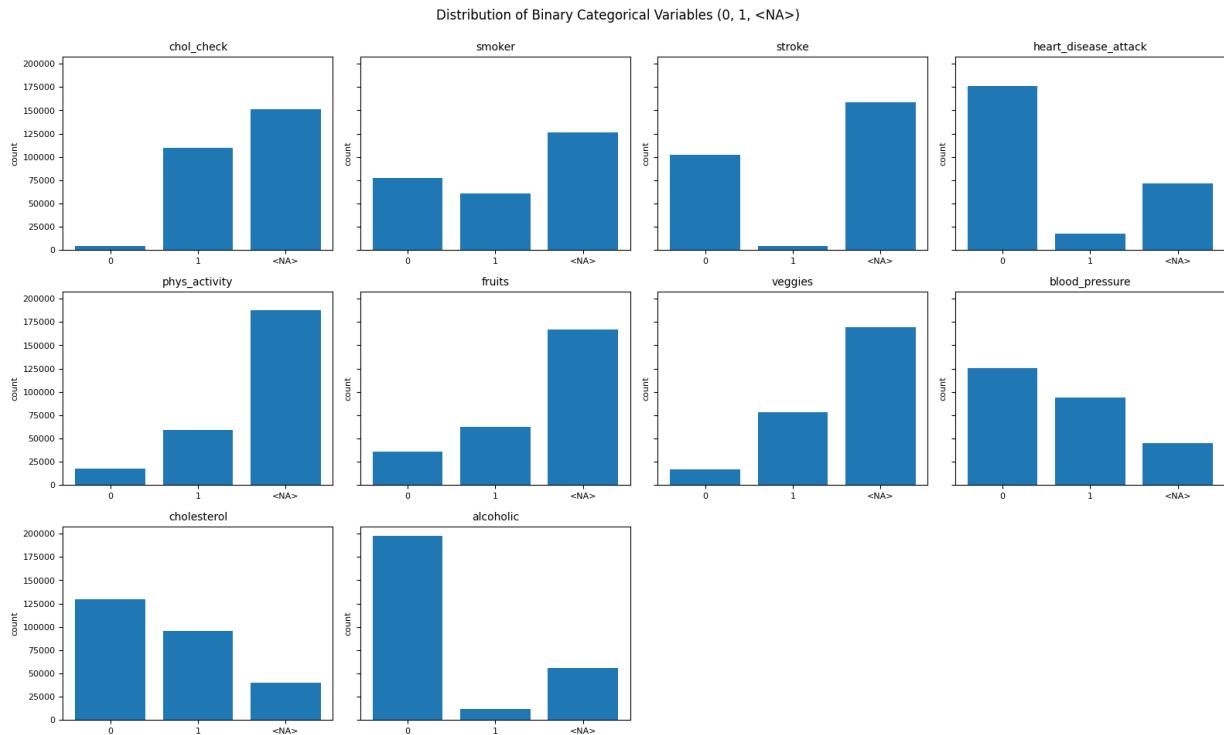


Figure 7: Distribution of Binary Categorical Variables (0, 1, <NA>)

For non-binary categorical variables, proportional mosaic-like plots were generated (Figure 8). These highlight the diversity of category levels across diabetes, education, general_health, and sex. Notably, education and general_health show relatively balanced category spreads, while diabetes is dominated by the "No" category.

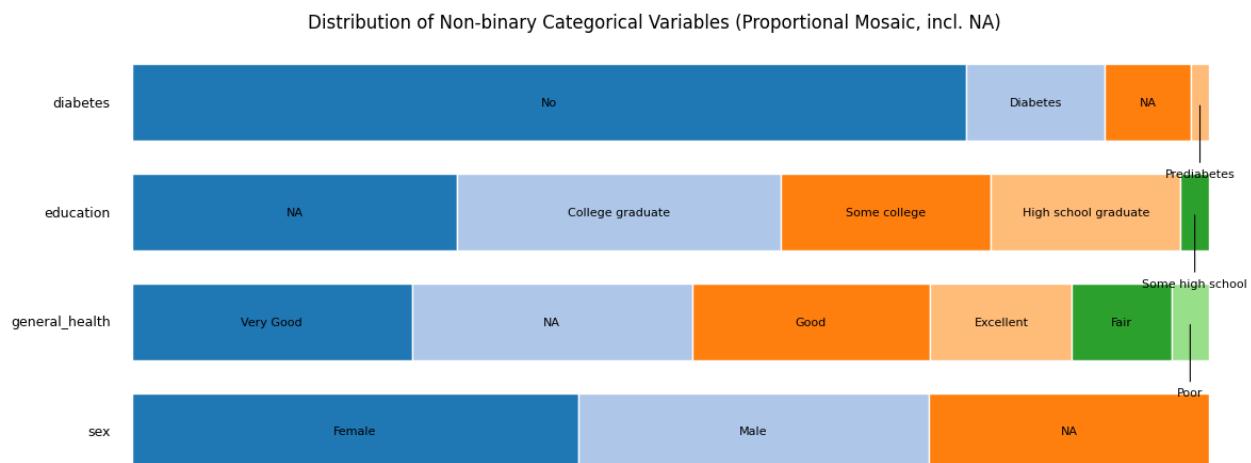


Figure 8: Distribution of Non-binary Categorical Variables (Proportional Mosaic)

Missing Value Handling – Categorical Variables

For the target variable (diabetes), rows with missing values were removed, resulting in the deletion of 21,184 records, which represented 8.0% of the dataset. This ensured that the target variable was complete for subsequent modeling.

For the binary categorical variables, a total of 10 features were processed using one-hot encoding with an explicit <NA> category. After encoding, the original binary columns were dropped, resulting in a clean set of dummy variables. Examination of the distributions showed that some features were strongly imbalanced (e.g., alcoholic dominated by 0), while others had more balanced patterns (blood_pressure, cholesterol). (The full distributional summary is provided in Appendix 2, Table 2.4.)

For the multi-class categorical variables (education, general_health, sex), missing values were explicitly recoded (e.g., "Unknown education", "Unknown health", "Not reported"). Categories with less than 5% frequency were merged into "Other". Specifically, "Some high school" in education and "Poor" in general_health were merged. The final category structures are visualized in (Figure 9), showing clearer and more interpretable groupings.

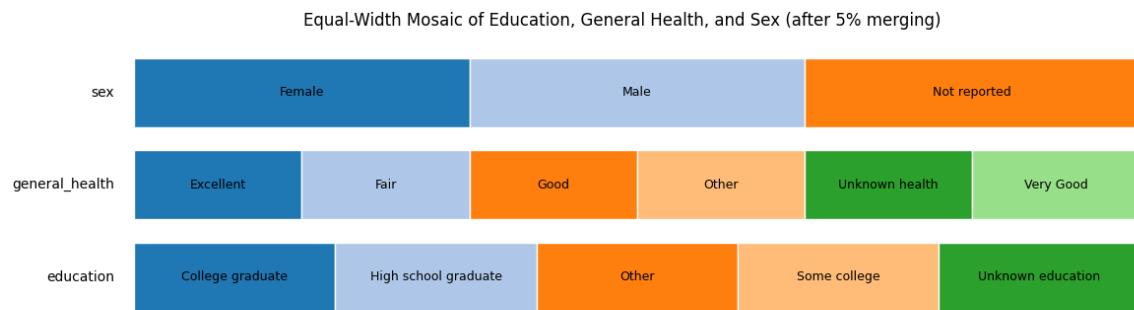


Figure 9: Final Distributions of Multi-class Categorical Variables (After Recoding and Merging)

Missing Value Handling – Numeric Variables

For the age and BMI variables, missing values were imputed using either the mean or median, depending on skewness: when $|skewness| > 1$, the median was used; otherwise, the mean. In addition, BMI was transformed with a Box-Cox or Yeo-Johnson transformation to correct skewness, since the variable is strictly positive.

For income, extreme outliers were reduced using Winsorization at the 1st and 99th percentiles. The variable was then transformed with Yeo-Johnson, missing values were imputed with the median in the transformed space, and the results were inverse-transformed back to the original scale.

The final distributions of the imputed and transformed variables are displayed in (Figure 10), showing improved normality for age, BMI, and income.

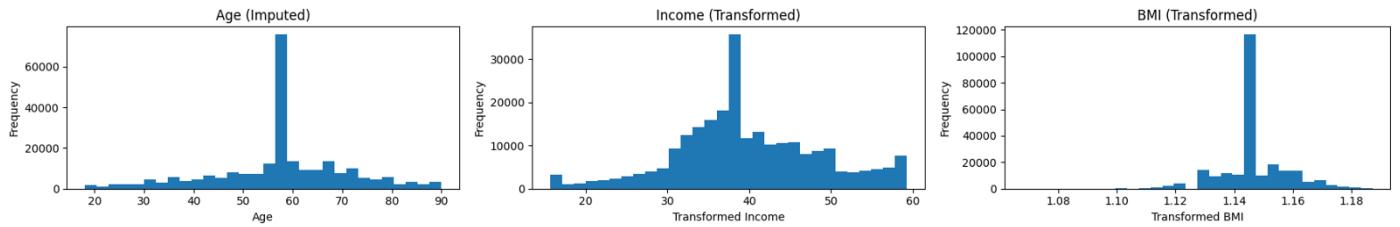


Figure 10: Distributions of Numeric Variables After Imputation and Transformation

Exploratory Data Analysis (EDA)

We inspect the cleaned dataset prior to modeling across four variable groups and a correlation view. The diabetes distribution shows marked class imbalance, with “No diabetes” dominating (Figure 11).

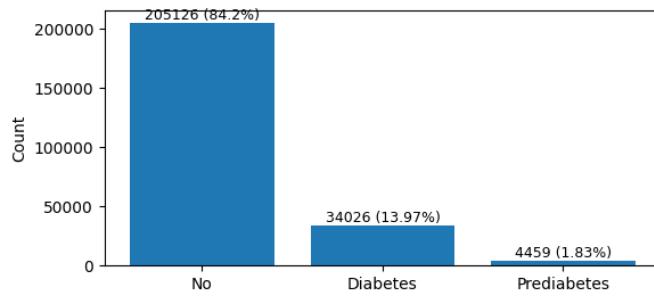


Figure 11: Diabetes distribution

For multi-class categorical variables (education, general_health, sex), proportional mosaics reveal heterogeneous category structures and the presence of explicit NA/Other levels (Figure 12). Detailed prevalence of binary features (rare vs. common) is summarized in the appendix (Appendix 2, Figure 2.3).

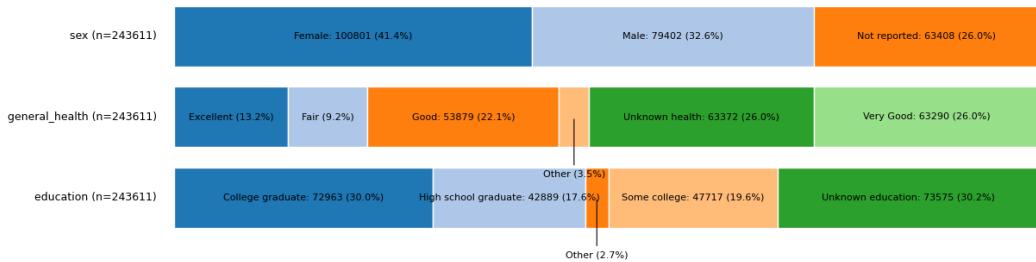


Figure 12: Multiclass distributions —Proportional mosaic

For numeric variables (age_imputed, bmi_imputed, income_imputed), histograms indicate residual skewness—particularly for income—despite prior transformations (Figure 13). On average, participants were middle-aged with mean BMI falling in the overweight range, while income showed high variability and extreme upper values. Distributional details and outliers are further illustrated via boxplots (Appendix 2, Figure 2.4), with target-stratified comparisons shown in (Appendix 2, Figure 2.5). Extended numeric descriptive statistics are reported (Appendix 2, Table 2.5).

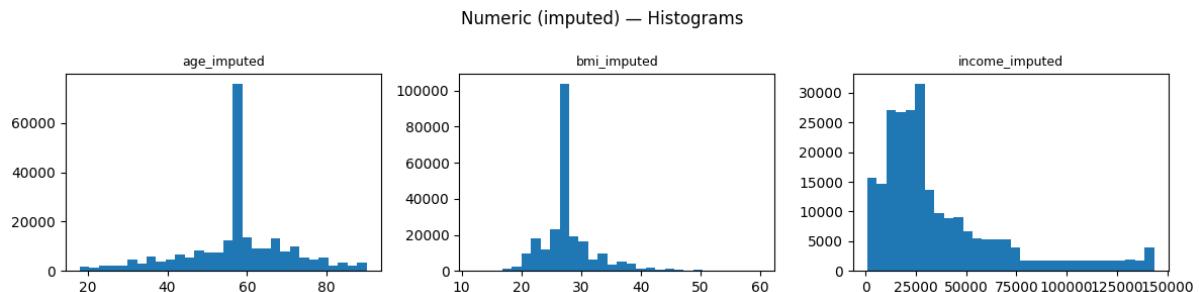


Figure 13: Numeric (imputed) — Histograms

The correlation matrix suggests generally weak associations among predictors, reducing multicollinearity concerns (Figure 14). The distribution of diabetes across binary features, highlighting uneven splits for some lifestyle/health factors, is provided for reference (Appendix 2, Figure 2.6).

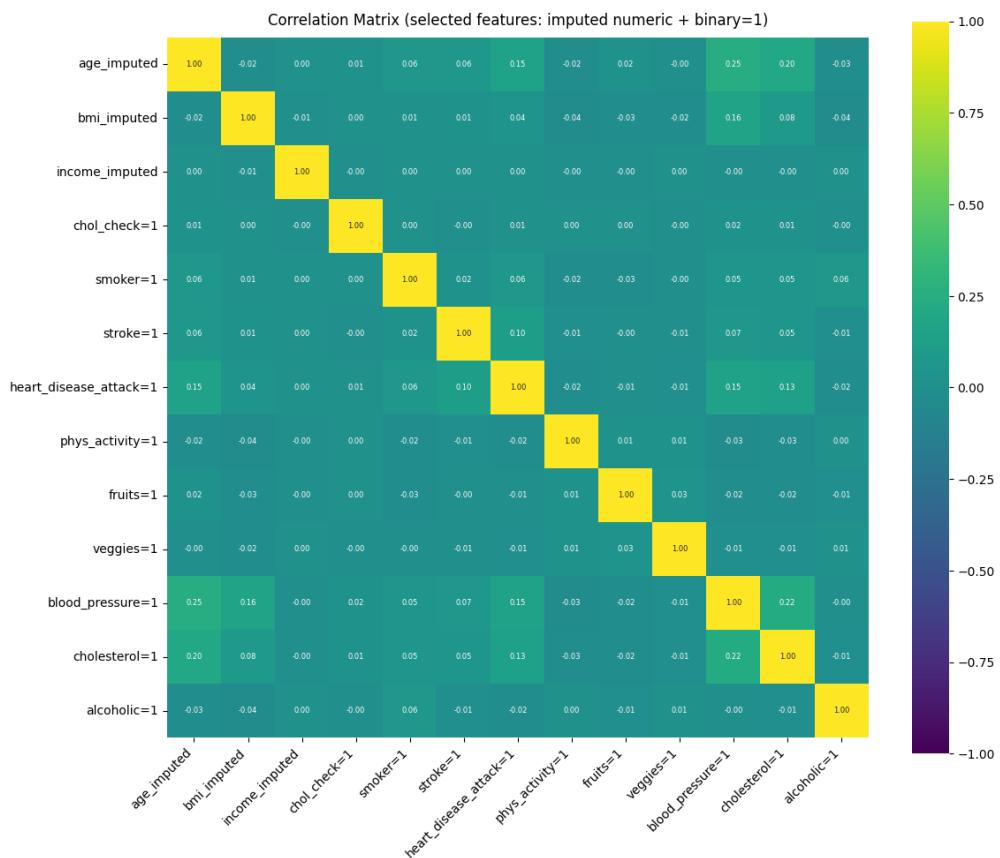


Figure 14: Correlation matrix (numeric imputed + binary = 1)

Cleaned Dataset Export

The cleaned dataset is exported as `cleaned_datasetB.csv` for further analysis.

Step 4. Exploratory Data Analysis

Numerical Features: Age and BMI

Two numerical features were selected for analysis: age (imputed) and BMI (transformed).

Distributions: Histograms show that age is centered around middle age, while BMI is right-skewed despite transformation (Figure 15).

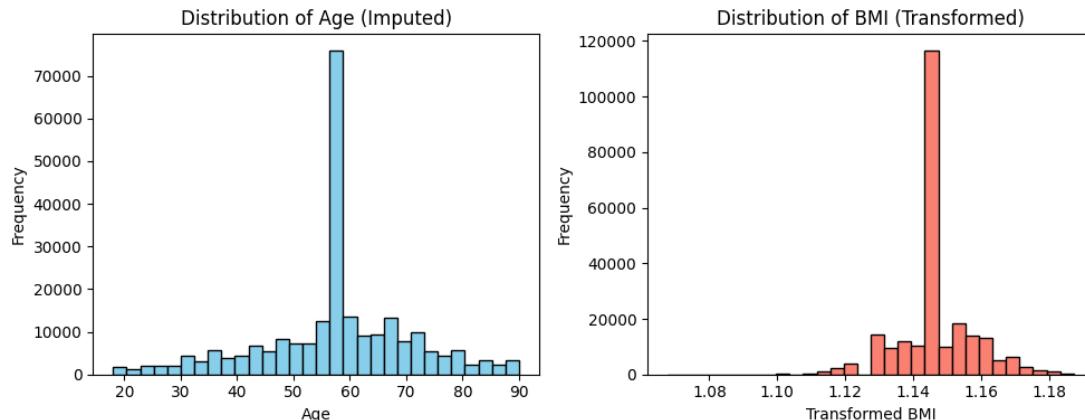


Figure 15: Distributions of age (imputed) and BMI (transformed)

Joint pattern: The scatterplot of age vs. BMI, stratified by diabetes status, suggests potential clustering but no strong separation between groups (Figure 16).

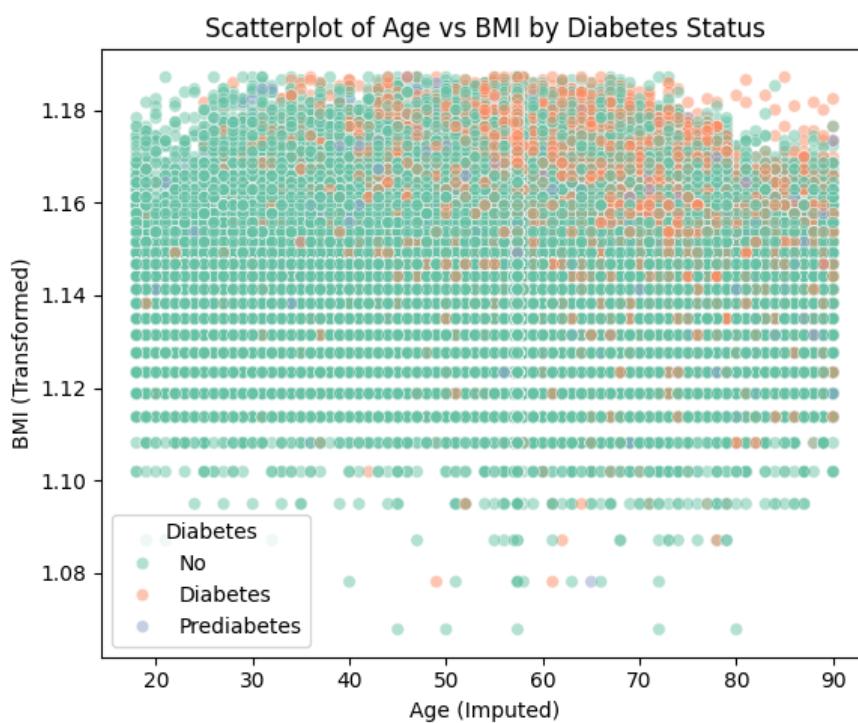


Figure 16: Scatterplot of age vs. BMI by diabetes status

Linear relationship: The correlation heatmap indicates a negligible linear association ($r \approx -0.01$) between age and BMI (Figure 17).

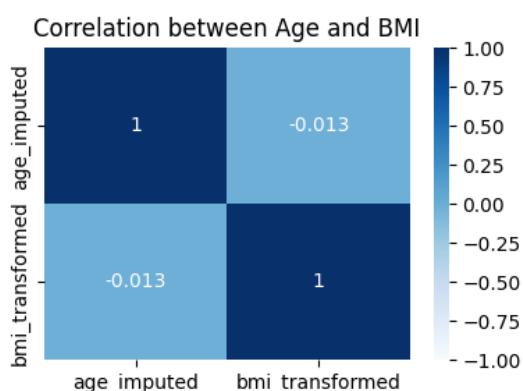


Figure 17: Correlation heatmap of age and BMI

Group comparison: A t-test confirmed that individuals with diabetes had a slightly higher mean BMI ($M = 1.15$, $SD = 0.01$, $n = 34,026$) compared to those without diabetes ($M = 1.15$, $SD = 0.01$, $n = 205,126$). The difference was statistically significant ($t = -89.74$, $p < 0.001$), with an effect size of Cohen's $d = 0.54$, indicating a moderate difference despite the small raw mean gap.

Summary: Age and BMI appear weakly related, and age does not differentiate diabetes status strongly. However, BMI shows a statistically significant and moderately sized association with diabetes, making it a more promising predictor than age.

Categorical Features: General Health and Sex

Two categorical variables were examined: General Health (Excellent, Very Good, Good, Fair, Other, Unknown health) and Sex (Male, Female, Not reported).

Bar plots (Figure 18) illustrate their overall distributions, while grouped bar plots (Figure 19) compare these variables across diabetes categories.

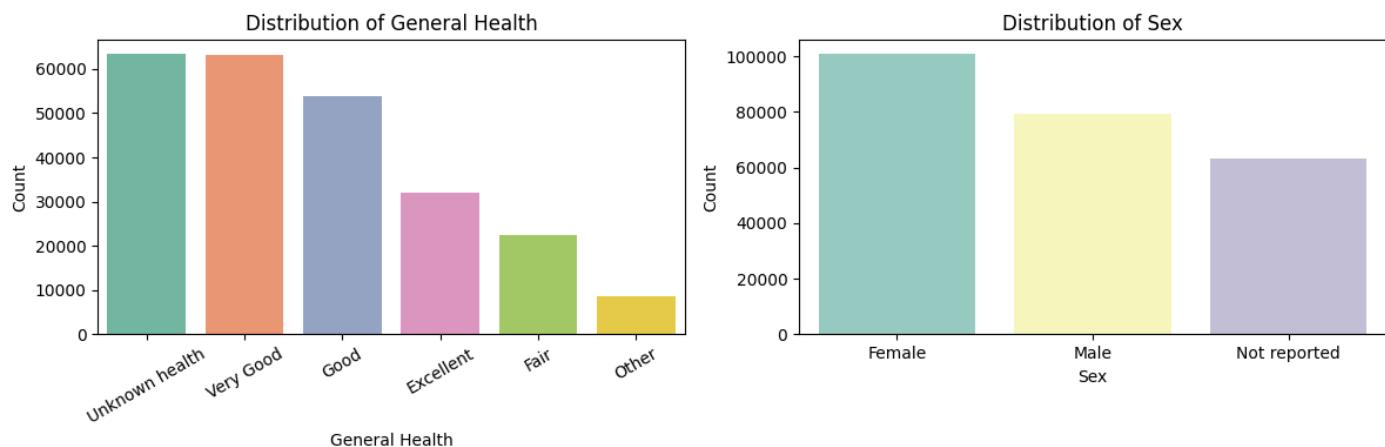


Figure 18: Distribution of General Health and Sex

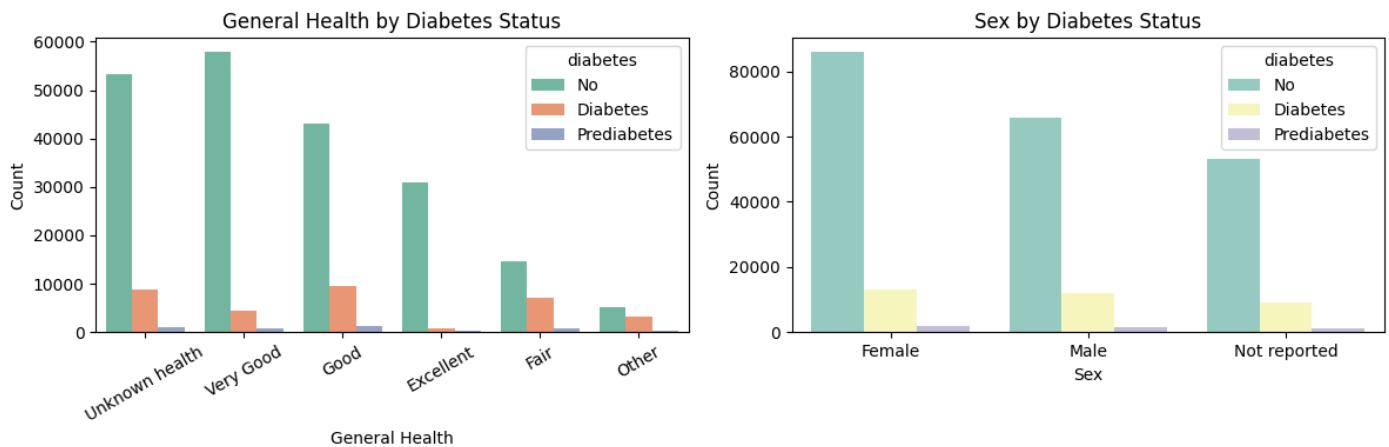


Figure 19: General Health and Sex by Diabetes Status

Chi-square tests confirmed strong associations between both features and diabetes status (General Health: $\chi^2 = 17284.10$, $p < 0.001$; Sex: $\chi^2 = 184.32$, $p < 0.001$). These results suggest that both self-reported health perception and sex distribution differ significantly between diabetic and non-diabetic individuals.

Summary: General Health shows a clear gradient where poorer health is more prevalent among diabetics, while sex differences are smaller in magnitude but remain statistically significant.

Summary of EDA Findings and Tests

This step integrates results from statistical analyses of the four selected features.

- BMI (transformed): Welch's t-test showed a significant difference between diabetic and non-diabetic groups, with a medium effect size (Cohen's $d = 0.54$).
- Age (imputed): Distribution was described using summary statistics (mean = 57.39, SD = 13.37); no formal test was conducted since age was mainly exploratory.
- General Health: Strong association with diabetes status was confirmed ($\chi^2 = 17284.10$, $df = 10$, $p < 0.001$).
- Sex: A weaker but still significant association was observed ($\chi^2 = 184.32$, $df = 4$, $p < 0.001$).

The consolidated findings are summarized in (Table 5).

	Feature	Type	Method	Statistic	p-value	Effect/Note	Conclusion
0	BMI (transformed)	Numerical	Two-sample t-test (Welch)	t = -89.74	0	Cohen's d = 0.54	Significant difference between diabetes groups
1	Age (imputed)	Numerical	Descriptive EDA	mean = 57.39	—	std = 13.37	Distribution characterized; no formal test rep...
2	General Health	Categorical	Chi-square test	$\chi^2 = 17284.10$	0	df = 10	Strong association with diabetes status
3	Sex	Categorical	Chi-square test	$\chi^2 = 184.32$	8.789e-39	df = 4	Weaker but significant association

Table 5: Summary of statistical tests and conclusions for selected features

Summary of Step 4.

The exploratory analysis of two numerical and two categorical features highlighted distinct patterns related to diabetes:

- Numerical variables: BMI showed a statistically significant difference between diabetic and non-diabetic groups (medium effect size), while Age mainly provided descriptive insights without strong group separation.
- Categorical variables: General Health exhibited a strong association with diabetes, indicating worse health status was more prevalent among diabetic individuals. Sex showed a weaker but statistically significant link.

Together, these findings suggest that both lifestyle/health indicators (e.g. BMI, general health) and demographic factors (e.g. sex, age) contribute to differentiating diabetes status, with BMI and general health being the most influential among the selected features.

Forest cover

Step 1. Problem definition

Problem

To what extent does soil type information improve the prediction of forest cover type compared to models relying solely on topographic and environmental variables?

Stakeholders and benefits

The findings of this study are relevant to multiple stakeholders engaged in forest management, land-use planning, and environmental conservation. Forestry managers and land-use planners stand to benefit directly, as improved predictive accuracy regarding forest cover can inform reforestation strategies and optimize the allocation of land resources. Environmental scientists and ecologists also gain from this research, since a clearer understanding of the relationship between soil type and vegetation patterns enhances ecological modelling, biodiversity conservation, and climate resilience assessments. At the policy level, government agencies responsible for environmental regulation and natural resource management can utilise these insights to design evidence-based policies, enforce sustainable land-use zoning, and balance economic development with ecological preservation. In addition, stakeholders within the forestry and agricultural industries may apply the results to reduce environmental risks, avoid land degradation, and increase the long-term productivity of their operations. Finally, local communities and Indigenous groups, who rely on forest ecosystems for cultural, economic, and ecological services, benefit indirectly from more sustainable forest management practices that safeguard ecosystem integrity and ensure intergenerational resource security.

Motivation

Understanding the factors that drive forest cover distribution is essential for sustainable land management, biodiversity conservation, and climate adaptation. While topographic and environmental variables are traditionally employed in predictive models, soil type is often overlooked despite its fundamental role in shaping vegetation growth and ecosystem dynamics. Without incorporating soil information, models risk underestimating the complexity of forest systems and may provide less reliable guidance for decision-making. The present study investigates the extent to which soil type improves the prediction of forest cover, resolving a critical knowledge gap that has direct implications for ecological research, forestry management, and policy development. A more accurate predictive framework serves to enhance scientific understanding of soil–vegetation interactions and to support practical applications.

Step 2. Data description

Table 1: Summary of dataset characteristic

Aspect	Description
Attributes and Instances	55 attributes and 30,860 instances
Data Types	45 are numerical (float), 9 are integer-valued, 1 is categorical [Appendix 3.1]
Missing Values	45 attributes contain missing values, accounting for approximately 26.2% of all entries in the dataset (Figure 1)
Unique Values	Forest Coverage contains 7 distinct categories: Spruce/Fir, Lodgepole Pine, Ponderosa Pine, Aspen, Douglas-fir, Krummholz, and Cottonwood/Willow [Appendix 3.2]

The numerical variables represent continuous environmental and geographical measurements, while the categorical variable Forest Coverage encodes the type of forest cover. The missing value highlights the need for careful handling of missingness before subsequent modelling.

Table 1 provides a concise summary of the dataset characteristics, including attribute counts, datatypes, missing values, and categorical uniqueness. To complement this tabular overview, Figures 1 illustrate the extent and distribution of missingness both across the entire dataset and per attribute. These visualizations highlight which features are most affected, offering additional context beyond the numerical summary.

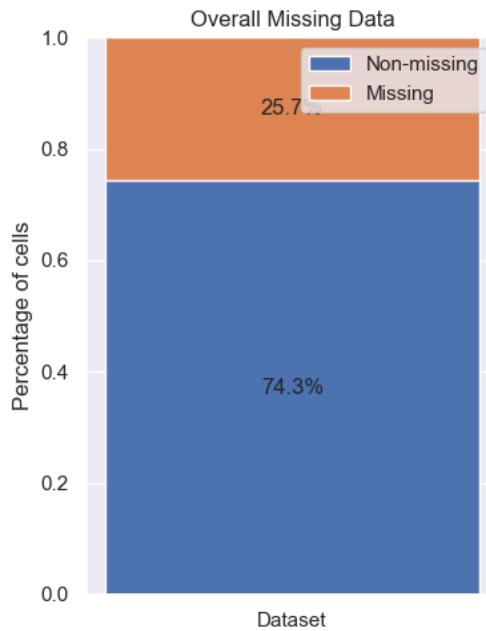


Figure 1: Overview of missing values in the dataset.

Step 3. Data cleaning and processing

Removal of Irrelevant Columns

The dataset was originally imported with an additional column "Unnamed: 0", a redundant index. Since this column carried no analytical meaning, it was removed in the dataset. This ensured a clean dataset structure without unnecessary identifiers.

Data Type Standardization

The dataset contains numerical variables stored in mixed numeric types (e.g., integers and floats), which may cause inconsistencies or errors during statistical analysis and modelling. Converting all numerical variables to a consistent float64 type ensures compatibility across analytical functions (e.g., correlation, imputation, model fitting) and prevents type-related errors, while preserving numeric precision.

Duplication of Records

The dataset may contain duplicate records, which can bias descriptive statistics and distort model training by over-representing certain observations. To address this issue, duplicate rows are removed. This cleaning strategy ensures that each observation contributes only once, thereby maintaining data integrity and preventing redundancy from inflating the dataset size or skewing analytical outcomes.

Consolidation of One-Hot Encoded Variables

Soil types and wilderness areas were originally represented as multiple one-hot encoded columns. This representation is redundant, inflates dimensionality, and makes interpretation more difficult without adding extra information. A cleaning strategy used is to collapse the one-hot encoded columns into single categorical variables. For soil types, all columns beginning with "Soil_Type" were identified and missing values replaced with zeros; if the sum of indicators in a row equalled one, the new variable was assigned "Known", otherwise "Unknown". For wilderness areas, the four columns (Neota, Rawah, Comanche Peak, Cache la Poudre) were combined such that if exactly one indicator equalled one, the corresponding area name was retained, otherwise the category "Other" was assigned. The original one-hot columns were then dropped. Collapsing the one-hot blocks into concise categorical variables improves interpretability and reduces redundancy, while still retaining the essential information needed for analysis. This also streamlines the dataset, ensuring that modelling focuses on meaningful categorical distinctions (e.g., Known vs Unknown soil information) without unnecessary dimensionality.

Detection and Treatment of Outliers

Invalid Entries

Some variables in the dataset contain values that fall outside of their logically valid ranges, such as negative slopes, hill shade values above 255, or aspects outside 0–360 degrees. To address this issue, valid ranges were explicitly defined for each variable and any entries lying below the minimum or above the maximum thresholds were recoded as missing values (NaN). This cleaning strategy ensures that invalid or impossible measurements do not distort descriptive statistics or bias model training, while preserving the integrity of the dataset by retaining only plausible observations.

Extreme Values

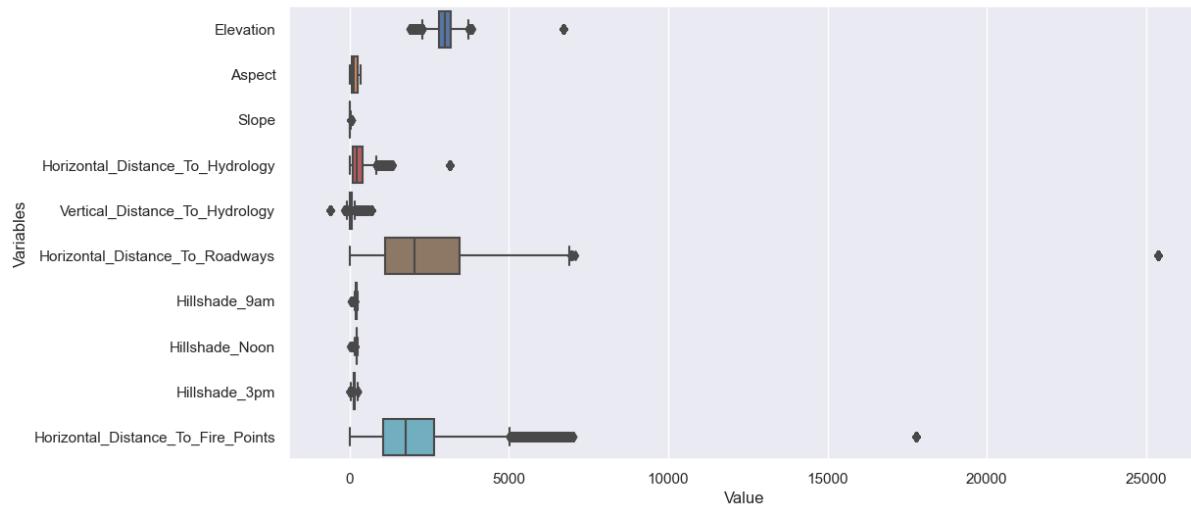


Figure 2: Parallel boxplots of numeric variables illustrating central tendency and extreme values

Figure 2 reveal the existence of extreme values in several numeric variables, including elevation and both horizontal and vertical distance measures. For example, negative values are observed in "Vertical_Distance_To_Hydrology", which are plausible because they indicate locations lying below the hydrological feature. "Horizontal_Distance_To_Roadways" and "Horizontal_Distance_To_Fire_Points" display very large range because there is no natural upper bound on how far a location can be from roads, hydrological features, or fire points. While these patterns highlight the presence of extreme cases, they do not all constitute invalid data. In this context, no corrective action was taken for high but plausible values, as tree-based models such as decision trees and random forests are relatively robust to outliers and can accommodate skewed distributions. Retaining the original values preserves the full variability in the dataset, ensuring that the models can learn from both common and atypical cases without introducing bias through artificial trimming.

Distribution of Categorical Variables

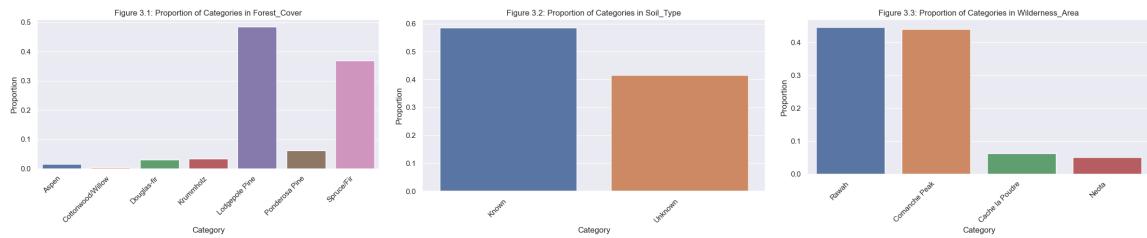


Figure 3: Distribution of categorical variables. Panel (a) shows the proportion of forest cover classes. Panel (b) illustrates the distribution of soil type information. Panel (c) presents the composition of wilderness areas.

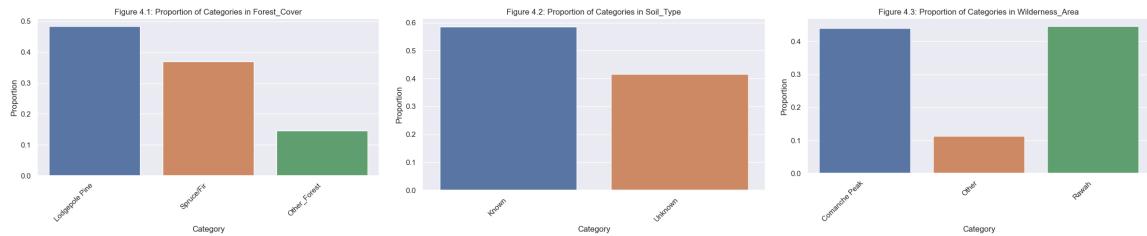


Figure 4: Distribution of categorical variables after cleaning. Panel (a) shows the proportion of forest cover classes. Panel (b) illustrates the distribution of soil type information. Panel (c) presents the composition of wilderness areas.

The initial distributions of the categorical variables are presented in Figure 3, which highlights a high degree of imbalance across categories. In particular, the "Forest_Cover" variable is dominated by Lodgepole Pine and Spruce/Fir, while several categories such as Aspen, Douglas-fir, and Cottonwood/Willow appear only rarely. Similarly, "Wilderness_Area" shows that Rawah and Comanche Peak account for the majority of observations, with Neota and Cache la Poudre contributing only a small fraction. To address this sparsity and reduce the risk of unstable model estimates, rare categories were consolidated into broader groups. As illustrated in Figure 4, the cleaned dataset retains only the two dominant forest cover classes explicitly, while the less frequent categories are grouped into "Other_Forest". For wilderness areas, the smallest classes were merged into "Other_Wilderness", producing a more balanced distribution across categories. This consolidation improves interpretability and ensures that categorical predictors provide meaningful signals in subsequent analysis without being distorted by sparsely populated classes.

Handling of Missing values

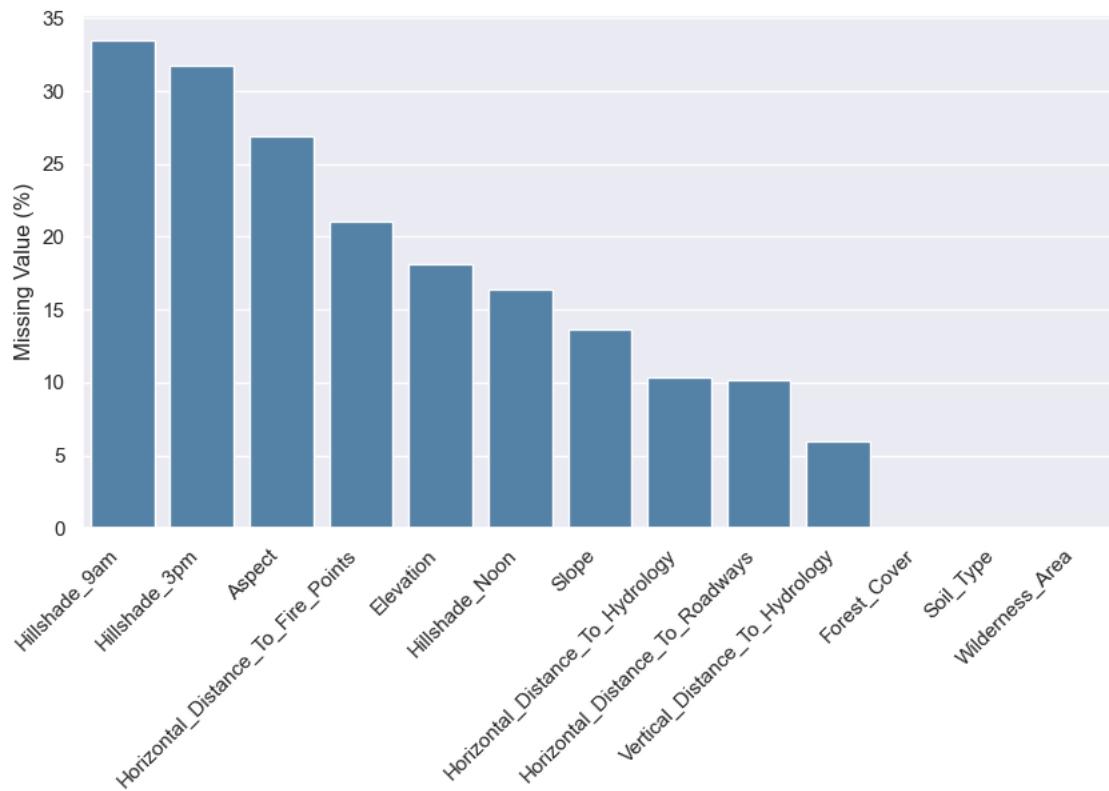


Figure 5: Missing value percentage by variable before handling

Hill shade values at 9am and 3pm exhibits substantial missingness over 30%, while noon had comparatively fewer missing values (Figure 5). Since noon often represents peak illumination, relying solely on it would bias the representation. Therefore we constructed “Hillshade_Mean_9_3” as the average of 9am and 3pm, capturing baseline morning-afternoon exposure (Figure 6). This separation of peak and baseline effects reduced dependence between variables and enhance interpretability.

For terrain variable such as elevation and slope, missing values were considered random measurement gaps without ecological meaning and were imputed using the median. Distances to hydrology and roadways were assumed to be universally measurable, so missing values were considered random errors and imputed with the median. Median imputation was adopted for continuous variables since most of them exhibits skewed distributions (Figure 2). In skewed data, the mean is sensitive to extreme values, whereas the median provides a more robust measure of central tendency.

Aspect was excluded from this process because it is a circular variable, its missingness was handled separately through sine-cosine transformation with an additional indicator variable. Since 0 and 360 represent the same direction, treating it as a linear numerical variable would be misleading. Therefore, we transformed aspect into sine and cosine components to preserve its directional properties while avoiding artificial discontinuities (Figure 6).

In contrast, missing values in “Horizontal_Distance_To_Fire_Points” were interpreted as meaningful, likely indication the absence of fire points nearby. Thus, they were retained with a binary indicator variable “FirePoint_Missing” (Figure 6). This prevents misleading imputations and retains the potential ecological meaning of missingness.

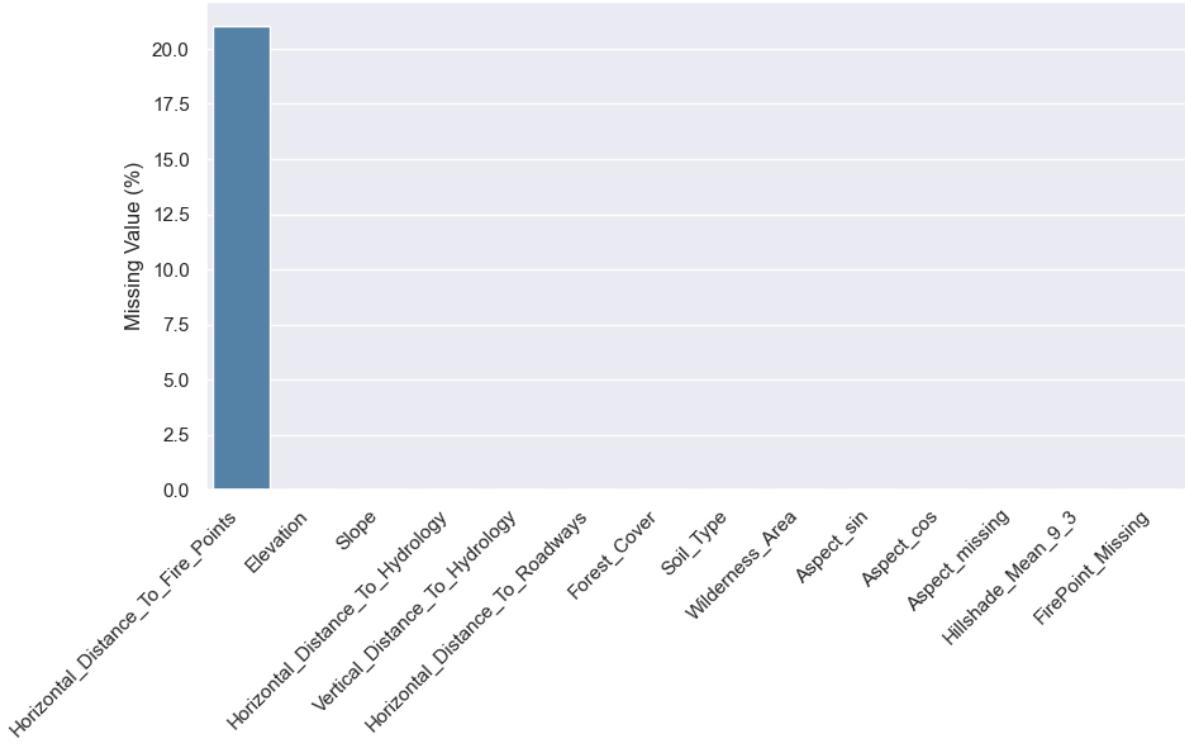


Figure 6: Missing value percentage by variable after handling

Independent and Dependent Variables

Attribute	Data Type
Forest_Cover	category

Table 2: data type for dependent variable

The dependent variable “Forest_Cover” was set as a categorical target, as shown in Table 2. This strategy enhances consistency and reduces the risk of type-related errors, while also ensuring that models can distinguish correctly between continuous predictors and categorical factors.

Attribute	Data Type
Elevation	float64
Slope	float64
Aspect_sin	float64
Aspect_cos	float64
Horizontal_Distance_To_Hydrology	float64
Vertical_Distance_To_Hydrology	float64
Horizontal_Distance_To_Roadways	float64
Horizontal_Distance_To_Fire_Points	float64
Hillshade_Mean_9_3	float64
Soil_Type	object
Wilderness_Area	category
Aspect_missing	float64
FirePoint_Missing	float64

Table 3: data types for independent variable

The dataset initially contained a mixture of numeric variables stored under inconsistent data types, such as integers and floats, alongside categorical variables represented in different formats (e.g., object vs. category). Such inconsistencies risk causing errors in statistical analysis and model training. To address this issue, all numeric variables were standardized to float64, ensuring uniform precision across continuous measures, while categorical predictors were explicitly cast into category type where possible. After these adjustments, the independent variables are summarized in Table 3, showing consistent float64 formatting for all continuous attributes, with "Wilderness_Area" properly encoded as a categorical variable.

Step 4. Exploratory Data Analysis

Correlation Analysis of Numerical Predictors

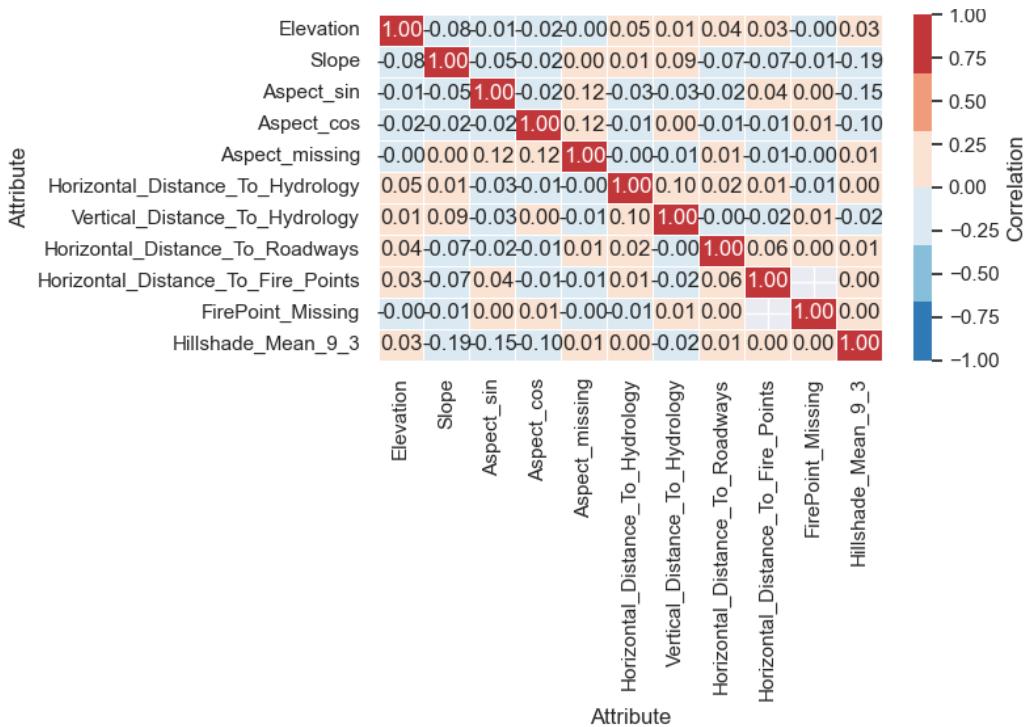


Figure 7: correlation heatmap

The correlation heatmap in **Figure 7** provides an overview of the relationships among the topographic and environmental predictors, which form the baseline set of features for comparison with soil type information. Most correlations are weak to moderate, with no strong multicollinearity observed (e.g., $|r| < 0.3$ for the majority of pairs), suggesting that the predictors contribute relatively independent information to the model. For example, slope and hillshade show mild negative correlation, reflecting terrain–illumination relationships, while distances to hydrology and roadways are largely uncorrelated with elevation. The choice of a correlation matrix provides a clear diagnostic of redundancy that could bias model training. The output shows that the baseline predictors are not highly collinear, meaning that any measurable improvement in predictive performance when soil type information is included can be attributed more directly to the added soil features, rather than being confounded by overlapping signal within the environmental variables.

Distributional Characteristic of Selected Numerical Variables

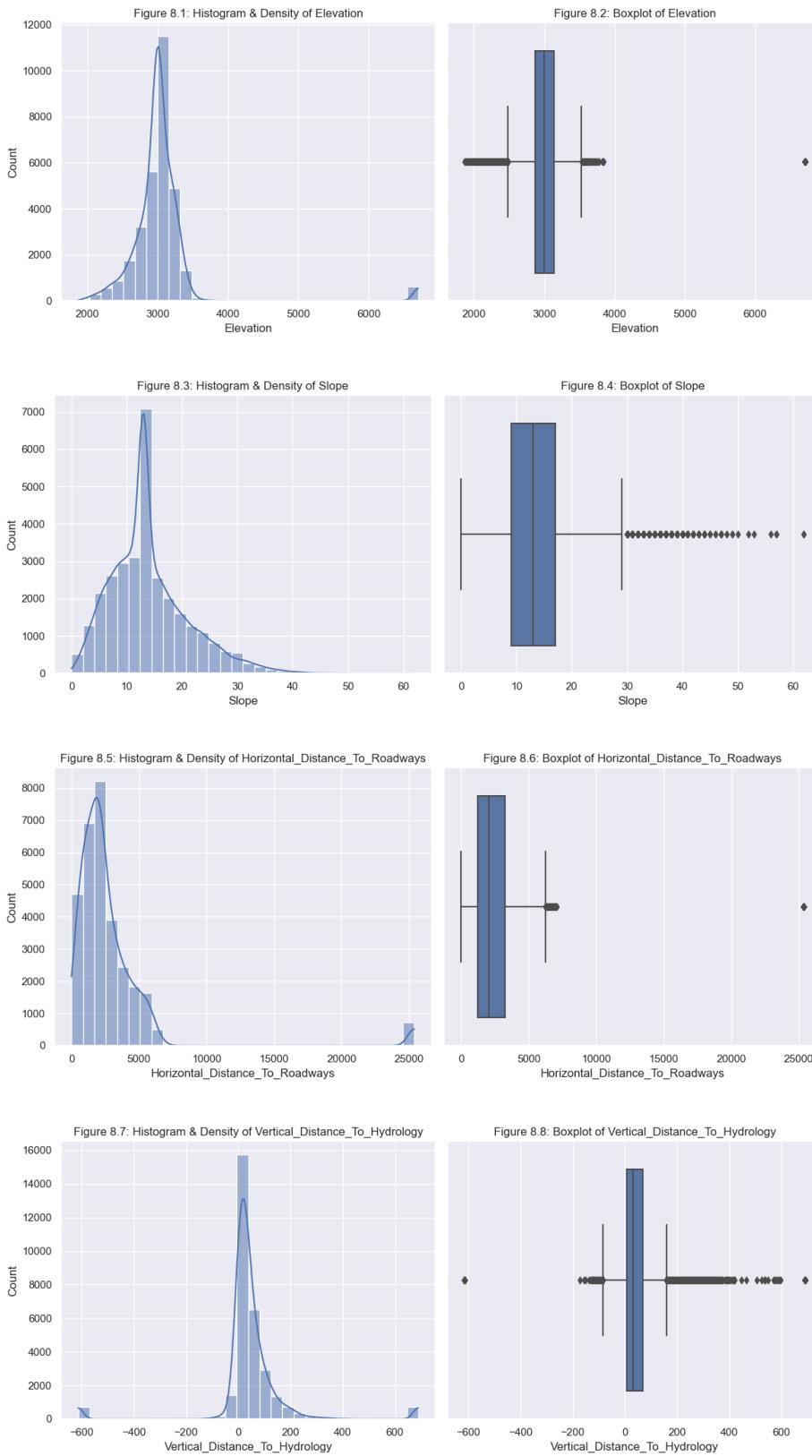


Figure 8: Distributional characteristics of selected numerical variables. Panels (8.1–8.2) distribution of Elevation. Panels (8.3–8.4) distribution of Slope. Panels (8.5–8.6) distribution of Horizontal_Distance_To_Roadways. Panels (8.7–8.8) distribution of Vertical_Distance_To_Hydrology.

The distributional characteristics of four representative numerical variables in the cleaned dataset are shown in Figure 8. A combination of histograms with density overlays and boxplots was chosen to provide complementary perspectives: histograms reveal the overall shape, skewness, and typical ranges of the variables, while boxplots emphasise central tendency and highlight the presence of extreme values.

This dual method is particularly important for environmental data, which often exhibit skewness and heavy tails: "Elevation" is approximately normally distributed around 3,000 meters, consistent with the study region, while slope and distance measures are highly right-skewed, with most values clustered near lower ranges but some extending to extreme outliers (Figure 8.1 – 8.6). Negative values in "Vertical_Distance_To_Hydrology" are retained as ecologically meaningful, representing sites located below nearby water features, whereas implausible negative elevations were already excluded during cleaning (Figure 8.7 – 8.8).

These results confirm that the dataset now captures both realistic central patterns and informative extremes. For future modelling, the presence of skewness and outliers supports the use of robust classifiers such as decision trees or random forests, which can incorporate both common and atypical conditions without being distorted by distributional assumptions. Importantly, establishing the statistical behaviour of the topographic and environmental variables provides a benchmark for assessing whether the inclusion of soil type information yields measurable improvements in predictive performance, thereby directly addressing the central research question.

Multicollinearity Diagnostic among Environmental Features

Table 4: Variance Inflation Factors (VIF) for environmental predictors.

	feature	VIF
0	Elevation	17.801850
10	Hillshade_Mean_9_3	16.155603
1	Slope	4.359205
8	Horizontal_Distance_To_Fire_Points	1.716026
7	Horizontal_Distance_To_Roadways	1.571888
5	Horizontal_Distance_To_Hydrology	1.512297
4	Aspect_missing	1.413458
2	Aspect_sin	1.169289
3	Aspect_cos	1.151331
6	Vertical_Distance_To_Hydrology	1.109099
9	FirePoint_Missing	nan

Multicollinearity among environmental predictors was assessed using the Variance Inflation Factor (VIF), as shown in Table 4. The results reveal that "Elevation" ($VIF \approx 17.8$) and "Hillshade_Mean_9_3" ($VIF \approx 16.2$) exceed the common threshold of 10, indicating a substantial degree of collinearity with other predictors, whereas all other variables fall within acceptable limits ($VIF < 5$). No corrective

action was taken at this stage because the planned predictive models—particularly tree-based methods—are less sensitive to collinearity compared to regression techniques. Nevertheless, documenting the presence of high VIF values is essential, as it underscores the importance of careful feature selection and provides a basis for later comparison when evaluating whether the inclusion of soil type information improves predictive performance.

Association Patterns of Selected Categorical Variables

Table 5: Crosstab of “Soil_Type” and “Forest_Cover” (row-wise proportions)

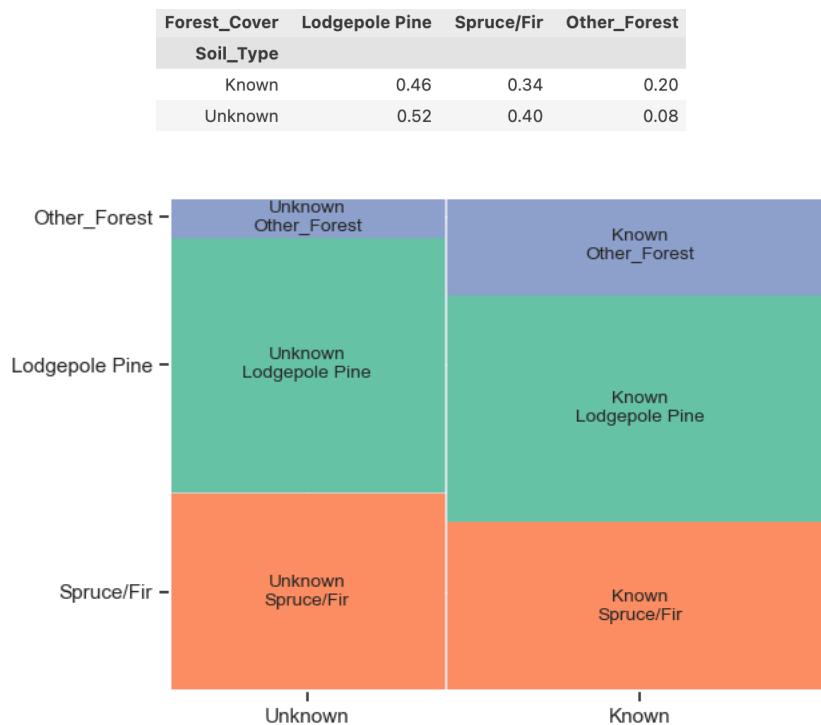


Figure 9: Mosaic plot of “Soil_Type” vs. “Forest_Cover”

The relationship between soil type and forest cover was examined using a cross-tabulation of proportions and visualised through a mosaic plot (Table 5, Figure 9). The issue investigated is whether soil information, even at the coarse level of “Known” versus “Unknown,” shows meaningful variation in forest cover composition. Row-wise proportions indicate that when soil type is Known, the forest cover distribution is more diverse, with Lodgepole Pine (46%), Spruce/Fir (34%), and Other_Forest (20%) all represented in substantial proportions. In contrast, when soil type is Unknown, the distribution becomes concentrated, with Lodgepole Pine (52%) and Spruce/Fir (40%) dominating, while Other_Forest declines sharply to just 8%.

The mosaic plot reinforces this pattern visually, showing broader category representation under known soil conditions and compressed diversity under unknown conditions. This method is justified because cross-tabulation highlights categorical associations numerically, while the mosaic plot provides an intuitive graphical summary of joint distributions. Together, these results suggest that the presence or absence of soil type information alters the balance of forest cover categories, offering preliminary evidence that soil information may contribute predictive value in modelling cover type, thereby directly linking to the research question.

Statistical Testing of Soil Type-Forest Cover Relationship

	Test	Chi ²	df	p-value
0	Soil_Type vs Forest_Cover	835.82	2	3.19e-182
1	Wilderness_Area vs Forest_Cover	6014.05	4	0

Table 6: Chi-square test result

The chi-square tests reported in Table 6 confirm that both "Soil_Type" and "Wilderness_Area" are strongly associated with forest cover type. For the soil variable, the test yielded a chi-squared statistic of 835.82 (df = 2, p < 0.001), indicating that the distribution of forest cover classes differs significantly between cases where soil type is known versus unknown. This aligns with earlier cross-tabulation and mosaic plot results, which showed a reduction in cover type diversity when soil information is missing. For wilderness area, the test statistic was even larger ($\chi^2 = 6014.05$, df = 4, p < 0.001), providing strong evidence that geographical location exerts a substantial influence on vegetation composition. These findings are important for the research question because they suggest that despite soil type being available only at a coarse binary level of known/unknown, it still contributes additional explanatory power alongside environmental and topographic variables. Incorporating soil type information may therefore improve predictive performance by capturing ecological variation not fully explained by terrain and distance measures alone.

Appendix

1.1

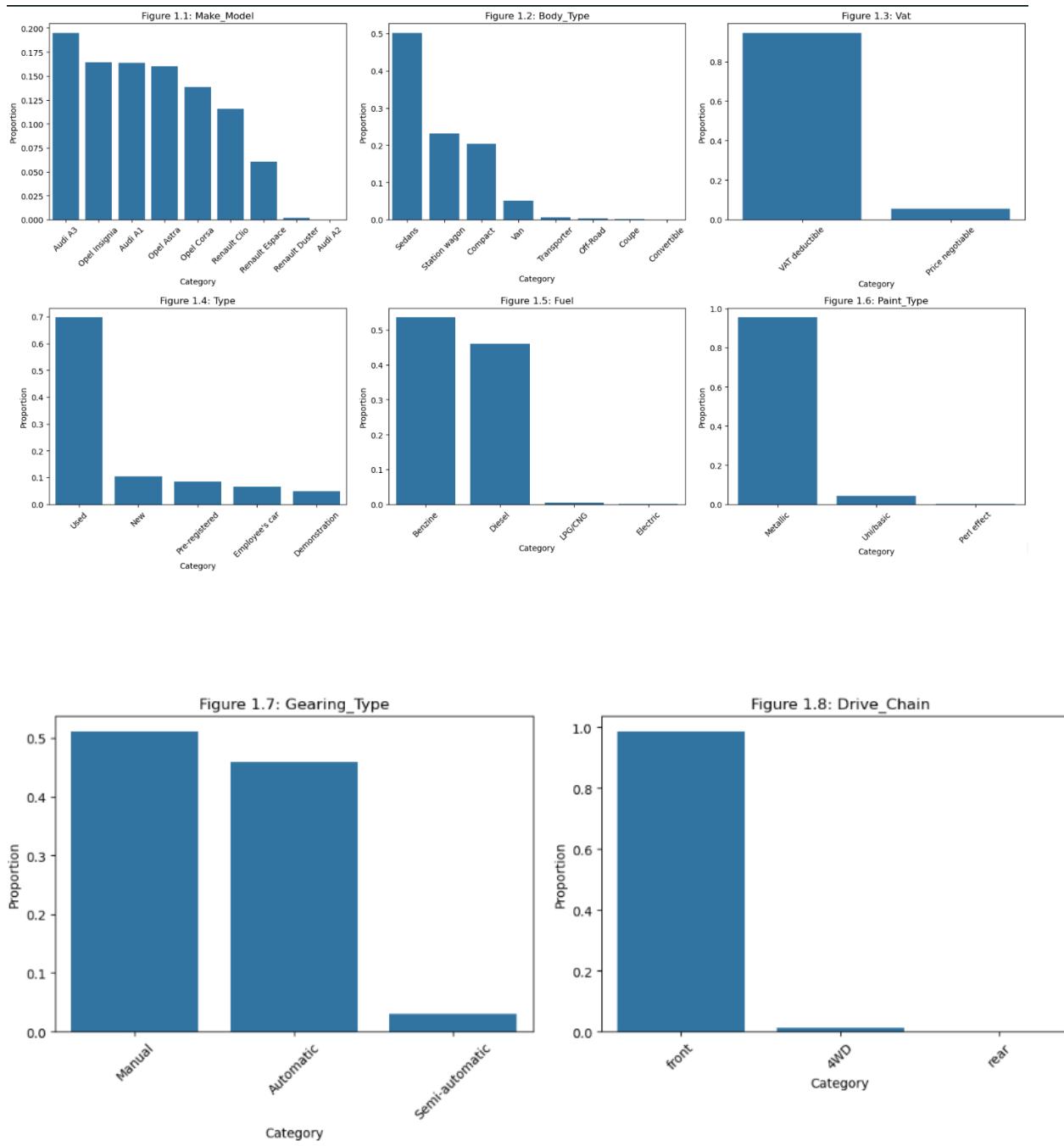


Table 2.1: Detailed Attribute Data Types and Categories

	Data Type	Attribute Category
Unnamed: 0	int64	Numerical
CholCheck	float64	Numerical
BMI	float64	Numerical
Smoker	float64	Numerical
Stroke	float64	Numerical
HeartDiseaseorAttack	float64	Numerical
PhysActivity	float64	Numerical
Fruits	float64	Numerical
Veggies	float64	Numerical
AnyHealthcare	float64	Numerical
NoDocbcCost	float64	Numerical
GeneralHealth	object	Categorical
Mental (days)	float64	Numerical
Physical (days)	float64	Numerical
DiffWalk	float64	Numerical
Sex	object	Categorical
Age	float64	Numerical
Education	object	Categorical
Income	object	Categorical
Diabetes	object	Categorical
BloodPressure	object	Categorical
Cholesterol	object	Categorical
Alcoholic	object	Categorical

Table 2.2: Missing Rates of All Attributes

	Attribute	Missing Rate (%)
0	Unnamed: 0	0.00%
1	CholCheck	57.00%
2	BMI	33.00%
3	Smoker	45.00%
4	Stroke	60.00%
5	HeartDiseaseorAttack	27.00%
6	PhysActivity	71.00%
7	Fruits	61.00%
8	Veggies	64.00%
9	AnyHealthcare	69.00%
10	NoDocbcCost	75.00%
11	GeneralHealth	26.00%
12	Mental (days)	49.00%
13	Physical (days)	54.00%
14	DiffWalk	54.00%
15	Sex	26.00%
16	Age	24.00%
17	Education	29.00%
18	Income	7.00%
19	Diabetes	8.00%
20	BloodPressure	17.00%
21	Cholesterol	15.00%
22	Alcoholic	21.00%

Figure 2.1: Distributions of Numerical Features (Horizontal Boxplots)

Distribution of Numerical Features (Horizontal Boxplots)

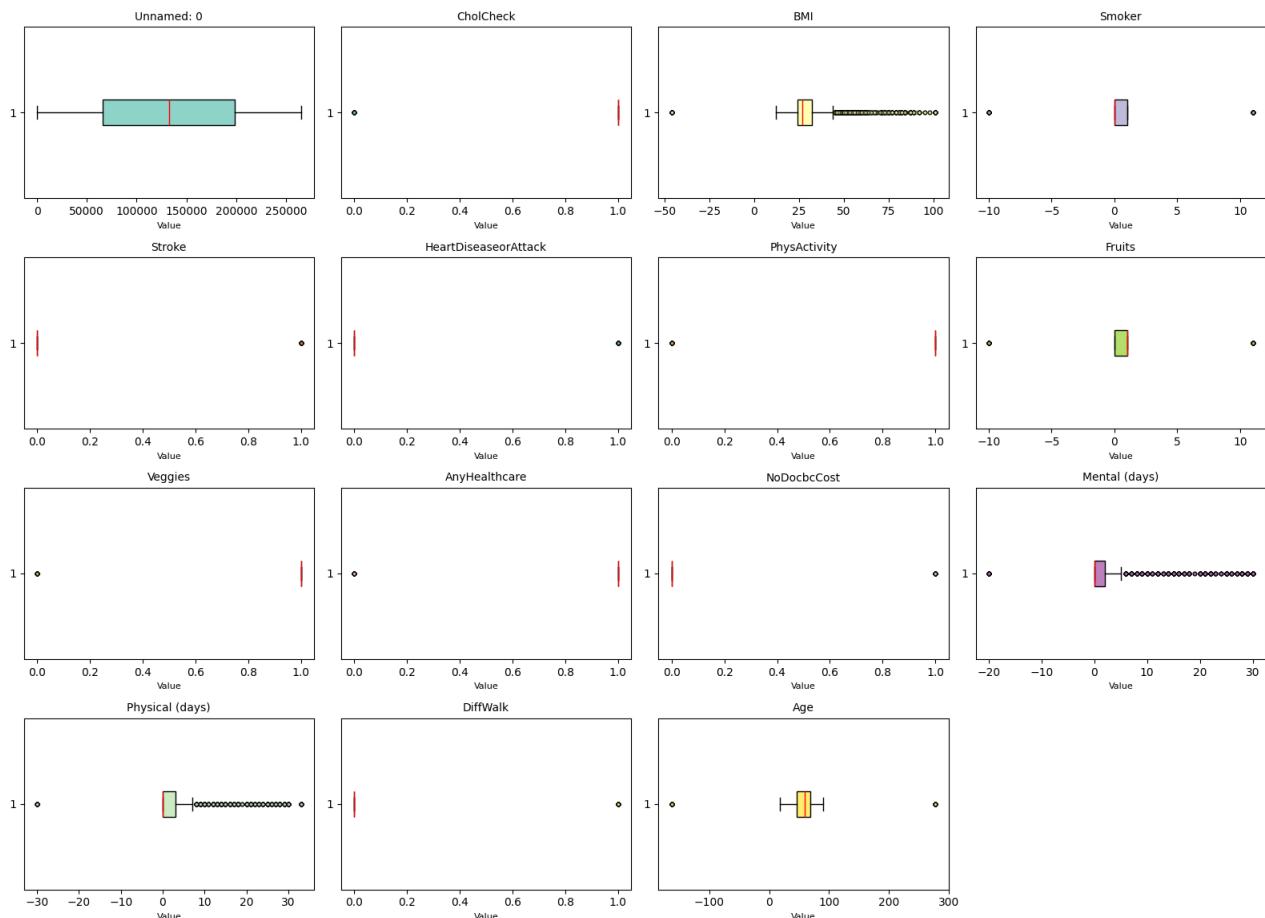


Table 2.3: Attribute Data Types Before vs After Formatting (Full Comparison)

Attribute Data Types Before vs After Formatting		
Attribute	Current Data Type	Original Data Type
chol_check	Int64	float64
bmi	float64	float64
smoker	Int64	float64
stroke	Int64	float64
heart_disease_attack	Int64	float64
phys_activity	Int64	float64
fruits	Int64	float64
veggies	Int64	float64
general_health	object	object
sex	object	object
age	float64	float64
education	object	object
income	float64	object
diabetes	object	object
blood_pressure	Int64	object
cholesterol	Int64	object
alcoholic	Int64	object

Table 2.4: Distributions of Binary Categorical Variables After One-Hot Encoding

	Variable	Level	Count	Percent
0	chol_check	0	3779	1.55%
1	chol_check	1	100870	41.41%
2	chol_check	NA	0	0.00%
3	smoker	0	71055	29.17%
4	smoker	1	56193	23.07%
5	smoker	NA	0	0.00%
6	stroke	0	93767	38.49%
7	stroke	1	3755	1.54%
8	stroke	NA	0	0.00%
9	heart_disease_attack	0	161867	66.44%
10	heart_disease_attack	1	15864	6.51%
11	heart_disease_attack	NA	0	0.00%
12	phys_activity	0	16317	6.70%
13	phys_activity	1	54435	22.35%
14	phys_activity	NA	0	0.00%
15	fruits	0	33094	13.58%
16	fruits	1	57208	23.48%
17	fruits	NA	0	0.00%
18	veggies	0	15730	6.46%
19	veggies	1	71955	29.54%
20	veggies	NA	0	0.00%
21	blood_pressure	0	115499	47.41%
22	blood_pressure	1	86652	35.57%
23	blood_pressure	NA	0	0.00%
24	cholesterol	0	119291	48.97%
25	cholesterol	1	87775	36.03%
26	cholesterol	NA	0	0.00%
27	alcoholic	0	181593	74.54%
28	alcoholic	1	10854	4.46%
29	alcoholic	NA	0	0.00%

Figure 2.2: Anomaly Heatmap of Binary Features

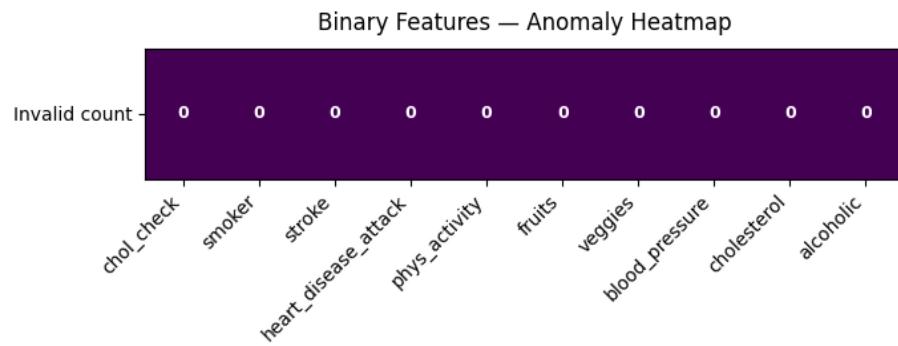


Table 2.5. Extended numeric descriptive statistics

	count	mean	std	min	1%	5%	25%	50%	75%	95%	99%	max
age_imputed	243611.0	57.386481	13.369668	18.0	22.0	32.0	52.0	57.386481	65.0	79.0	88.0	90.0
bmi_imputed	243611.0	27.784579	4.841297	12.0	19.0	21.0	26.0	27.0	29.0	37.0	46.0	60.0
income_imputed	243611.0	37150.614807	32670.008139	827.0	878.0	4353.0	15747.0	25041.0	47752.0	117358.0	143086.7	143583.28

Figure 2.3. Binary categorical variables — counts of 1 vs. 0

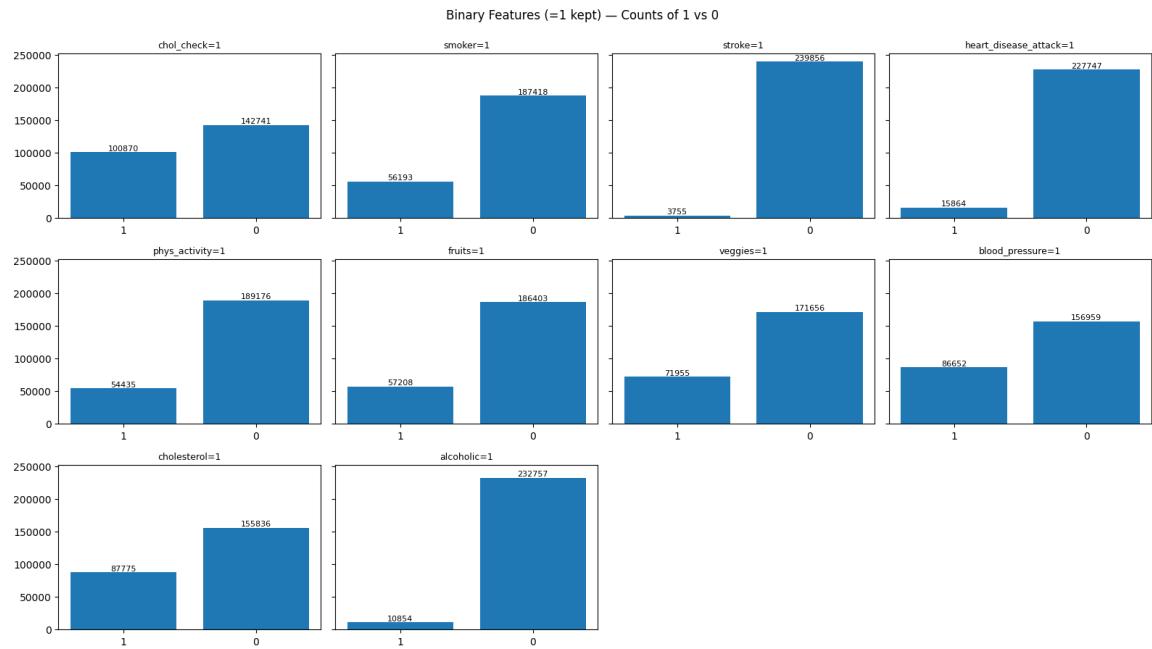


Figure 2.4. Numeric (imputed) — Boxplots

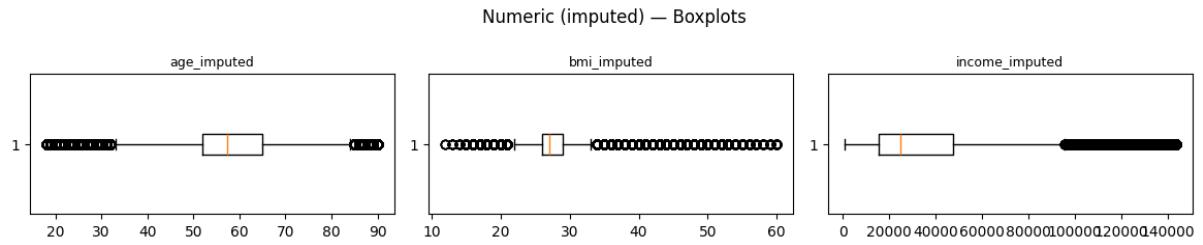


Figure 2.5. Numeric variables by diabetes status (boxplots)

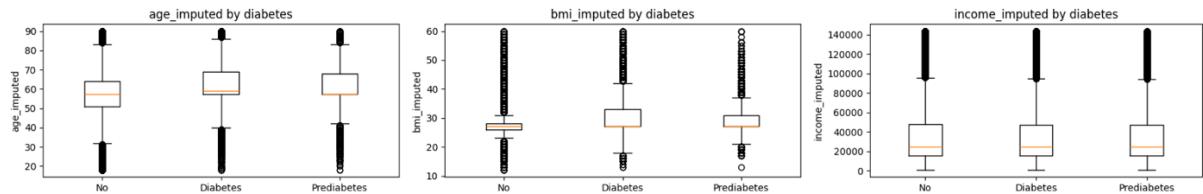
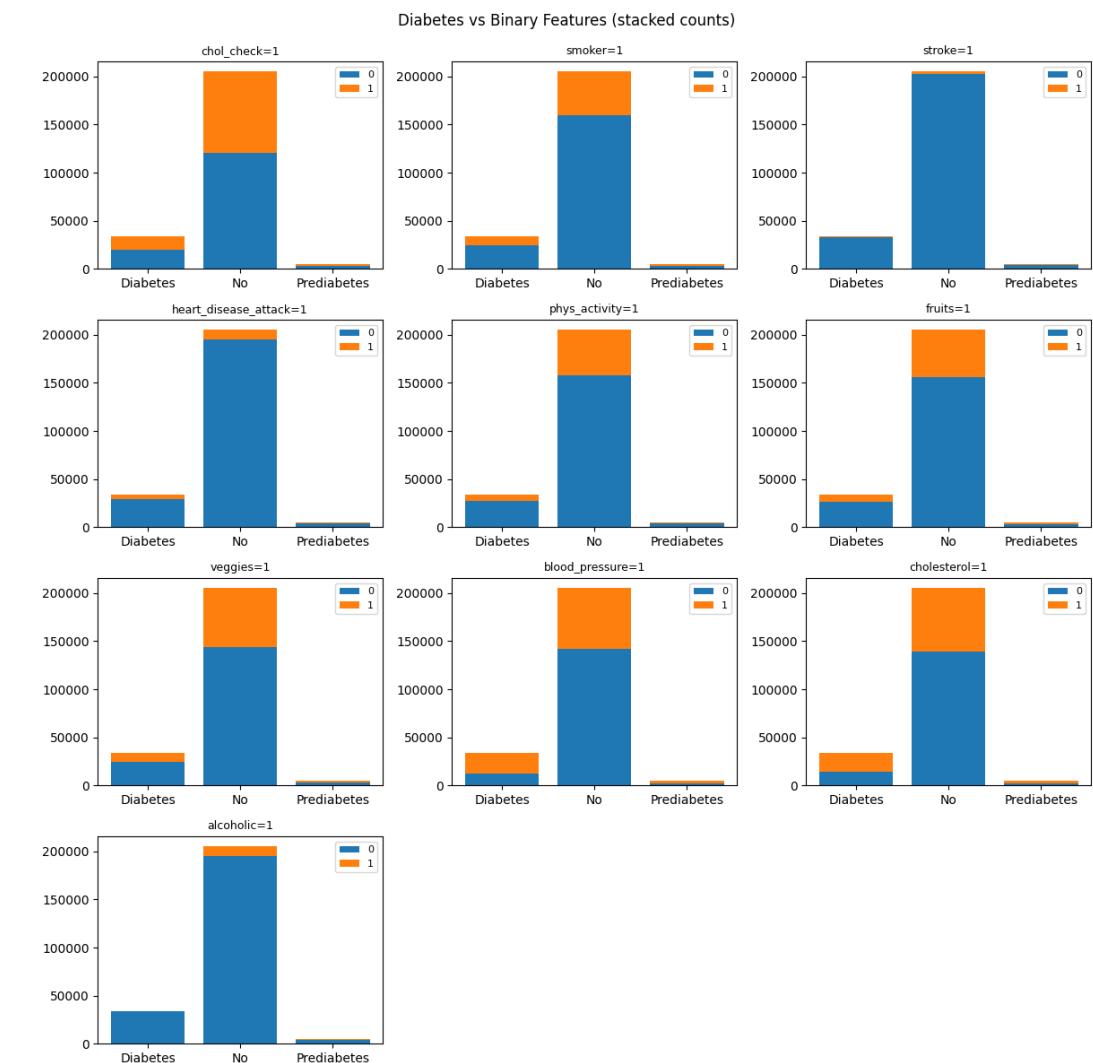


Figure 2.6. Diabetes vs. binary features (stacked bar plots)



3.1 Data types of forest Cover Attributes

Attribute	Data Type
Unnamed: 0	int64
Elevation	float64
Aspect	float64
Slope	float64
Horizontal_Distance_To_Hydrology	float64
Vertical_Distance_To_Hydrology	float64
Horizontal_Distance_To_Roadways	float64
Hillshade_9am	float64
Hillshade_Noon	float64
Hillshade_3pm	float64
Horizontal_Distance_To_Fire_Points	float64
Soil_Type1	float64
Soil_Type2	float64
Soil_Type3	float64
Soil_Type4	float64
Soil_Type5	float64
Soil_Type6	float64
Soil_Type7	int64
Soil_Type8	int64
Soil_Type9	int64
Soil_Type10	int64
Soil_Type11	int64
Soil_Type12	float64
Soil_Type13	float64
Soil_Type14	float64
Soil_Type15	float64
Soil_Type16	float64
Soil_Type17	float64
Soil_Type18	float64
Soil_Type19	float64
Soil_Type20	float64
Soil_Type21	float64
Soil_Type22	float64
Soil_Type23	float64
Soil_Type24	float64
Soil_Type25	float64
Soil_Type26	float64
Soil_Type27	float64
Soil_Type28	float64
Soil_Type29	float64
Soil_Type30	float64
Soil_Type31	float64
Soil_Type32	float64
Soil_Type33	float64
Soil_Type34	float64
Soil_Type35	float64
Soil_Type36	float64
Soil_Type37	float64
Soil_Type38	float64
Soil_Type39	float64
Soil_Type40	float64
Forest_Cover	object
Neota	int64
Rawah	int64
Comanche Peak	int64

3.2 Summary of Categorical Variables

Variable	Unique Categories	No. of Categories
0 Forest_Cover	['Spruce/Fir', 'Lodgepole Pine', 'Ponderosa Pine', 'Aspen', 'Douglas-fir', 'Krummholz', 'Cottonwood/Willow']	7