Project Stage 1 – Group 46

Group Member:

490051481 lshe0103

540650820 zxia0226

540237964 ywan0205

Auto Prices

UNIKEY: ywan0205

1.Topic and research question

1.1 Problem Description

All these cast a shadow on the growing demand for affordable cars. But determining a fair price for a used car can be difficult because the cost depends on many factors, including the brand, model, age, mileage, fuel type and additional amenities, such as air conditioning or navigation equipment. Traditional pricing methods commonly work on seller experience or simple formulas that lead to impractical prices. That can mean buyers end up paying more than necessary, and sellers sell their cars for less than they're worth. Consequently, the marketplace is imbalanced and transparent pricing is difficult to sustain.

1.2 Business and Research Need

With increasing competition in the used car market, possessing a system with the ability to anticipated pricing is becoming even more critical. Buyers want to be sure they're paying a fair price, while sellers and dealerships want to price things to attract buyers and maximize profit. To win the trust of users and promote sales, online car platforms must also provide feasible price recommendations. Similarly, insurers and automotive finance companies rely on accurate pricing to evaluate risks and offer appropriate premiums or loan amounts. So why we built data-driven price prediction model, so everyone could take smarter decisions, and the market becomes more transparent and efficient.

1.3 Research Question

"What are the most important attributes that affect vehicle prices?"

1.4 Benefits of Solving the Research Question

Research question can benefit used vehicles buyers by providing them more convenience through a better idea of factors that affect more price of used vehicles. Dealerships and online marketplaces can also use findings from this research to adapt pricing strategies and inventory management. For instance, a recent study of the automotive market indicated that data-driven pricing models can be used to mitigate pricing errors through early detection and prevention.

2.Data description

2.1 Dataset Overview

This project works on auto_price dataset. csv, and includes detailed data about used cars, including widespread prices, specifications or features used by car owners, and many more. The data consists of both numerical and categorical data like Price_Value, Mileage_Value, Age, Make_Model, Body_Type, Fuel. The data has been processed for missing values, units conversions, and key variables normalization.

2.2 Number of Attributes and Instances

The dataset contains 32 attributes (columns) and 18,286 instances (rows).

2.3 Data Dictionary

A comprehensive data dictionary is provided in Appendix 1.

3.Data ingestion and cleaning

3.1 Data ingestion

Python Pandas has been used to import and handle the dataset for preprocessing and analysis. In our project, we load the data by calling pd.read_csv('auto_price.csv'), which converts the raw CSV file into a Pandas DataFrame. This DataFrame is structured in a tabular format where each row represents a used car and each column corresponds to a specific feature (e.g., Make_Model, Body_Type, Price, Mileage, Fuel, etc.).

3.2. Data quality assurance and cleaning

In this project, I ensured data quality by carefully ingesting and cleaning the dataset using Python Pandas. I began by importing the data with pd.read_csv('auto_price.csv'), which loaded the CSV file into a DataFrame. I then examined the data's structure using df.shape, df.info(), and df.head() to identify any quality issues. I noticed that there were missing values in both numerical columns (such as "Gears", "Age", "Previous_Owners", "Inspection_New", and "Cons_Comb") and categorical columns (such as "Make_Model", "Body_Type", "Fuel", etc.).

To address these issues, I filled missing values in numerical columns with the median using df[col].fillna(df[col].median()), as the median is less sensitive to outliers and helps preserve the overall distribution. For categorical columns, I filled missing values with the mode by using df[col].fillna(df[col].mode()[0]), which maintains the most frequent category in the data. I also checked for duplicate rows using df.duplicated() and removed them with df.drop_duplicates() to avoid redundant information. Additionally, I dropped the "Unnamed: 0" column (generated during import as an index) using df.drop('Unnamed: 0', axis=1) since it did not contribute any useful information.

To ensure consistency across key features, I created custom functions like standardize_price(), standardize_mileage(), and standardize_weight(). These functions cleaned the data by removing currency symbols, converting units (such as converting miles to kilometers and pounds to kilograms), and formatting the values appropriately.

Overall, I used a combination of Pandas functions (e.g., pd.read_csv(), df.fillna(), df.drop_duplicates(), and df.drop()) along with custom Python functions to clean and standardize the dataset. This rigorous data cleaning process was crucial to improving data quality and ensuring the reliability of the subsequent machine learning models.

4.Exploratory data analysis (EDA)

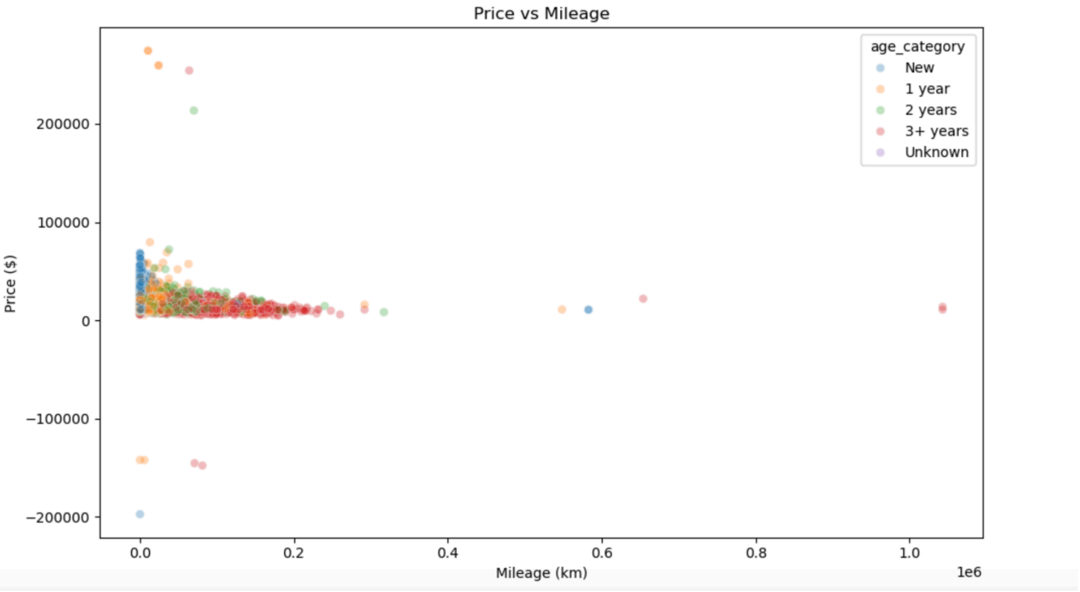


Figure 1: Scatterplot of Price vs. Mileage

This scatterplot shows a correlation of Price_Value and Mileage_Value. In this visualization, I also represented different age categories by using different colors. Main takeaways from this figure are:

1.There is a noticeable negative correlation in mileage and price meaning that vehicles with higher vehicle mileage are priced lower.

2.A newer vehicle — by year category — will command a higher price than an older one within a comparable mileage range.

Implications for Modeling:

Such patterns suggest that mileage and vehicle age are important attributes that influence price. These variables should be included as main and interaction terms, based on their priority in the modeling phase. Driven by their positive correlation with price, this indicates that the relation between these variables and price is strong, thus suggest careful feature engineering around these variables can improve the model performance.

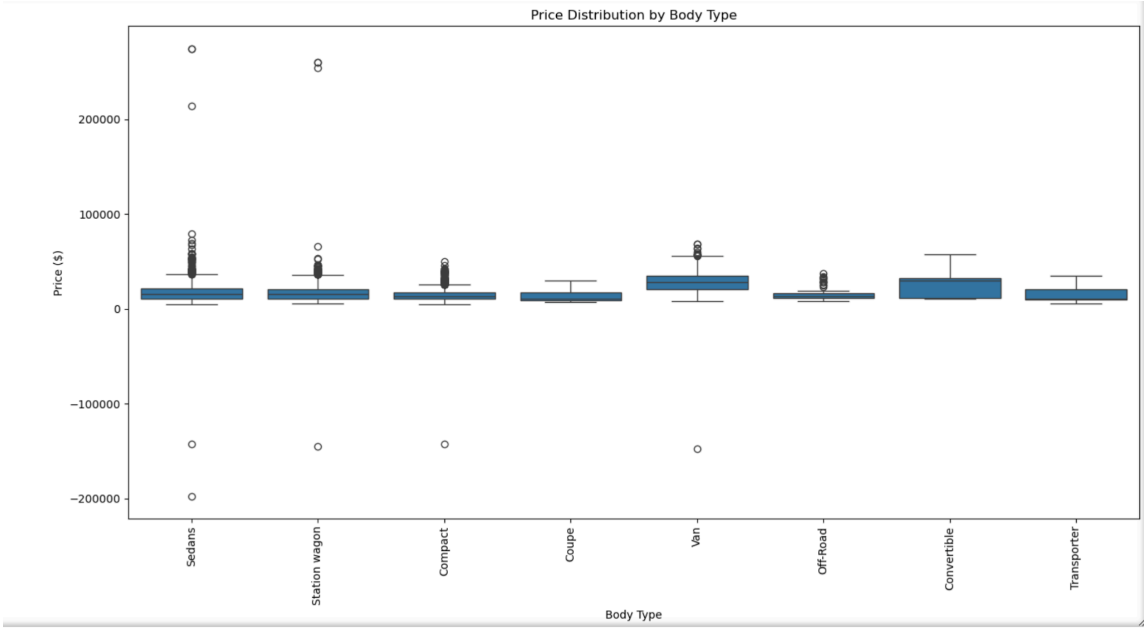


Figure 2- Boxplot of Price Distribution by Body Type

Let us take a look at this boxplot which compares the price distribution among vehicle body types. This figure provides box plots of the median, interquartile ranges and outlier for each body type category. The key observations are:

1.For example, SUVs and luxury sedans have a much higher median price and a much broader price range than more common body types like compact or hatchback vehicles.

2.The range of price within each category suggests that vehicle design is an important factor in determining price.

Implications for Modeling:

This analysis demonstrates that the body type represents an important categorical variable on vehicle prices. When creating predictive models, body type is a necessity feature that should be included probably through one-hot encoding or something similar. Knowing this distribution is very useful because it allows to define a reasonable threshold and treat outliers accordingly — all of which helps to create a more robust model.

In summary, this finding indicates that the body type is among the categorical variables that have a significant effect on the price of vehicles. So when creating predictive models we should include body type as a feature (e.g. through one-hot encoding or similar methods) so we can properly calculate its impact.

Diabetes Diagnosis

UNIKEY: zxia0226

1. Topic and research question

1.1 Background and Research Motivation

With the changes in people's modern lifestyles, including shifts in dietary habits, decreased physical activity, and increased work and life pressures, diabetes has become one of the major public health issues worldwide. Many traditional screening tools focus on single physical indicators such as blood pressure or weight, which mainly reflect physiological conditions, while ignoring potential social and behavioral factors that could also contribute to diabetes—such as mental health, mobility, education level, and income. With the increasing ways of investigating, we could gain abundant and various data and combine them to analyze the role nontraditional factors play in the prediction of diabetes, thereby doing some favor for public health intervening and personalizing medicine strategies.

1.2 Research Question

Is it possible to effectively predict diabetes using lifestyle and mental health characteristics without relying on blood glucose-related indicators?

1.3 Business and Research Need

This research have significant meaning: for individuals, some normal people don’t have money or time to do regular and accurate diagnosis, if it is possible to predict diabetes by just check out their lifestyle and mental health, people could earlier do some precaution and take the treatment from doctors or adjust some of their daily habits; for governments and policymakers, high-risk populations can be identified more easily and target interventions accordingly with better resource allocation; for healthcare providers, clinicians could do some initial judgments if laboratory data are not immediately available.

2. Data description

2.1 Dataset overview

The dataset I have chosen is ‘diabetes_diagnosis.csv’, this is a dataset with abundant data, which may be derived from a large healthy survey. It has 264,802 records and 22 attributes, every record represents an individual respondent’s information related to health status, behaviors, and sociodemographic characteristics. The key attribute in this dataset is ‘Diabetes’, a binary variable which represents whether the respondent has been diagnosed with diabetes or not(1 for Yes, 0 for No). Other attributes cover a variety of aspects, including health behaviors(e.g., smoking and alcohol use), chronic diseases and functional status(e.g., stroke, heart attack and difficulty walking), subjective health and psychological status and social demographic characteristics(e.g.,sex, age group, education level and income level). This dataset contains a variety of data types including integer, float and strings. However, since this dataset is a raw collection, most of its attributes have missing values, may also contain invalid or inconsistent records, so it is necessary to do some data cleaning for the missing data while maintaining core features to provide a reliable dataset for further research.

2.2 Data Dictionary

A comprehensive data dictionary is provided in Appendix 2.

3.Data ingestion and cleaning

3.1Data ingestion

The dataset this research uses is ‘diabetes_diagnosis.csv’, I imported it via Python’s pandas library then stored it as a DataFrame structure by using the pandas.read.csv() function. The dataset has 264,802 records and 23 attributes. More data details have already been explained in the Data description.

3.2 Data quality assurance and cleaning

To make sure that the dataset is accurate and dependable, a comprehensive data cleaning process was performed. I ran a comprehensive data cleaning with the Python Pandas library. The dataset was loaded using pd.read_csv(), and its properties and quality were examined using functions such as df.info(), df.describe(), and df.isnull().sum(). Several attributes such as PhysActivity, Fruits, and AnyHealthcare, have over 70% missing values. They are all removed by df.drop() so as not to allow for any disturbances or harmful biases that might result from analysis in these cases. For key features with moderate missingness such as BMI, Smoker, GeneralHealth, and Mental (days), data is filled according to the appropriate strategies: continuous variables like mean imputation should be done to reduce sensitivity to outliers and categorical variable fill using mode. I also made data types of different columns consistent: including Sex, Cholesterol, BloodPressure into object type categorical fields; and made conversion automatic for binary variables that need it into int64. After cleaning, the dataset contains 119,260 records and 17 attributes and each data type is well defined with no remaining missing value. All features had well-defined and consistent data types, including 7 categorical variables and 5 numeric fields. These steps were necessary to maintain data integrity, eliminate potential sources of bias, and ensure that the dataset was ready for reliable statistical analysis and machine learning modeling.

4. Exploratory Data Analysis (EDA)

As for whether lifestyle and mental health features can be used to predict diabetes without relying on such traditional biomedical indicators as the level of sugar in one's blood, I have conducted a thorough Exploratory Data Analysis (EDA). This stage of the research is not modeled. The two visualizations below were selected because of their relevance to answering our research question--as well as clearly displaying patterns that could influence predictive modeling.

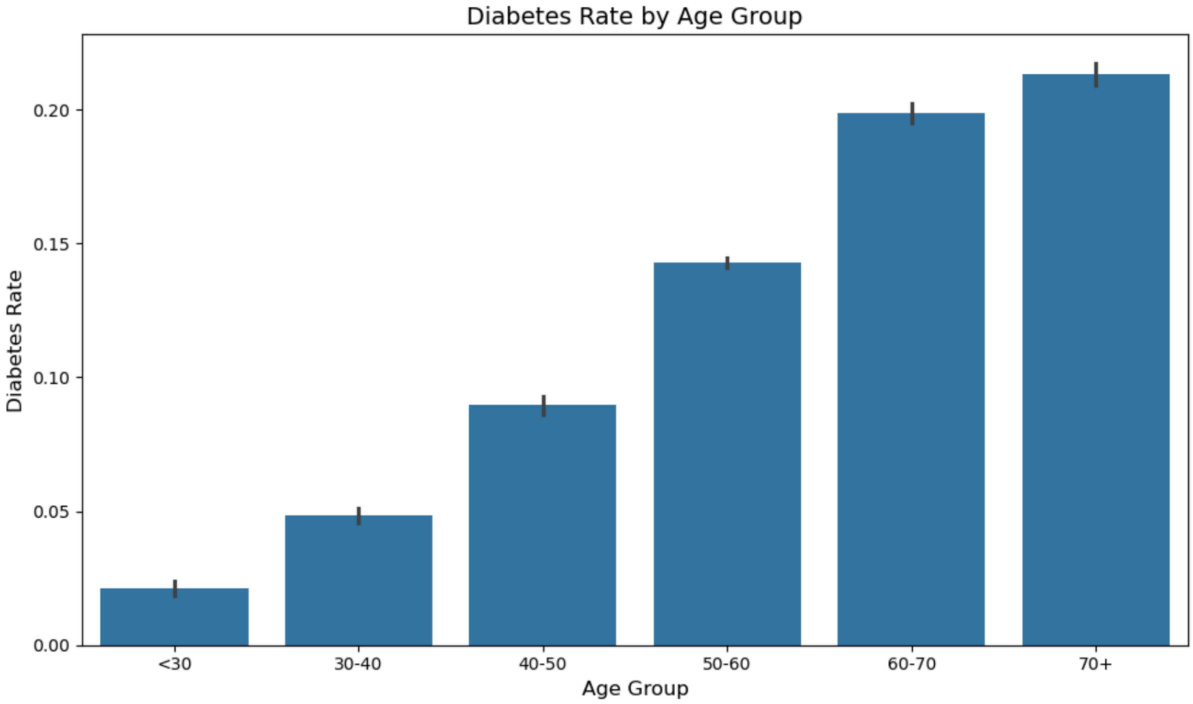


figure1

As for the first figure, a histogram presents the proportion of people diagnosed with diabetes by age. As the elderly population continues to grow, both urban and rural areas should recognize this trend and work promptly to prevent it. The diabetes rate has a strong positive correlation with the age of the subject. For those beyond 70, it reaches approximately 23%. But those under 30 years old are only 1.5 %. It goes up smoothly through middle age without peaks and then slows at 46 – 47%. The change is neither linear nor uniform with advancing age; after 50 it speeds up. This indicates that growing older itself is a considerable risk factor. Although not itself a behavioral variable, age is indicative of many noninvasive variables that accumulate over time--for example accumulated poor diet, financial insecurity or a long period spent sedentary. From the standpoint of public health, this trend emphasizes the need for targeted prevention strategies to start at an earlier age. Like many other studies of a quasi-experimental design, it is confirmed here that age can build a very powerful predictive model which does not rely on laboratory data. Besides, segregation across age brackets demonstrates that age may well interact in meaningful ways with various lifestyle variables. So further model development language interaction terms including, for example, age-lifestyle combinations will make this approach a useful guide to better health practices.

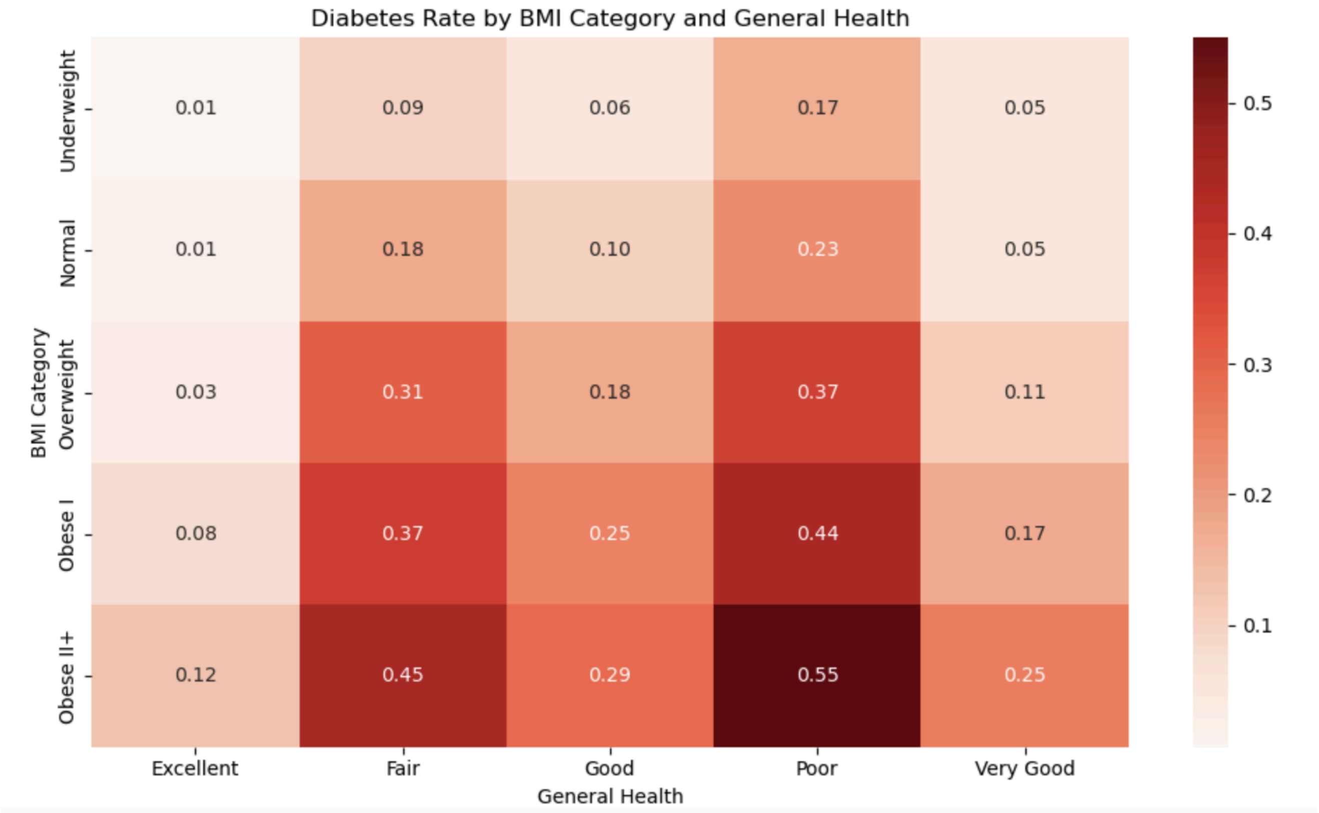


figure2

Figure2 is a heat map showing diabetes prevalence in various combinations of BMI groups and self-reported general health. It appears that the general health level and physical condition have a complicated interaction and relationship. People who are classified as Obese II+ and report poor general health have the highest diabetes rate, reaching 59%; those in the same BMI category but report excellent health have 13%+ lower rates. Similarly, for respondents with a normal BMI, diabetes rates increase sharply from a mere 1% among the excellent health group to 25% in those reporting poor health. The common theme is that perceived general health is an important risk factor for Individual Diabetes, combined with BMI or independently. The tendency observed in each plane corresponds beautifully to this idea, that the perceived health profile encompasses all sorts of potential risk factors—the stress, fatigue and undiscovered underlying diseases which may go unnoticed simply because they are not expressed as high Body Mass Index (BMI). With both these questions being non-invasive, possible to answer through surveys, the second graphic lends added weight to using lifestyle and mental health/feeling data as reliably predictive indicators in non-clinical diabetes risk models.

Overall, these charts show convincingly there are some non-medical characteristics which are providing consistently strong and clear signals in predicting diabetes. For instance, age is the sum of all risk factors accumulated over time--it's long-term background; while BMI represents livesystle in attitude toward life and quality of health. General health perception matters in part because it speaks to each participant's body. And more critically, none of these three measures requires laboratory testing. They·can~ all be collected through self-reported surveys, wearable technology or simple screening tools.Such findings support the central hypothesis of this paper: Diabetics can indeed be predicted with insights from lifestyle and mental health, without having to undertake direct blood glucose measurements. Furthermore, this analysis highlights the importance of careful feature selection and preprocessing for any subsequent modelling work. Among the variables considered in this study we have added non-linear relationships between them as well as looking at interactions in relation to used features.

Forest_Cover

UNIKEY: lshe0103

1. Topic and research question

Forests play a crucial role in maintaining ecological balance by providing habitats for wildlife, regulating climate, and preventing soil erosion. Understanding the factors that influence forest cover is essential for managing ecosystems and supporting biodiversity. In this research project, the research question is: **"How does forest cover depend on the relationship between soil type and elevation?"** This study investigates whether certain tree species are primarily suited to specific soil types or if elevation also plays a significant role in determining forest cover.

Understanding this relationship is critical for predicting which tree species will thrive in particular regions, thus contributing to more effective ecological restoration and conservation efforts. This research holds significant value for both environmental protection agencies and botanists. For environmental agencies, the findings can inform tree planting strategies, helping to maximize survival rates by selecting species that are best adapted to local soil conditions. For botanists, this study deepens our understanding of tree species distribution across different wilderness areas, enhancing our knowledge of plant ecology. Additionally, by analyzing soil types, botanists may be able to predict which tree species might be found in previously unexplored regions, contributing to the discovery of new biodiversity.

2. Data description

The dataset contains data collected from four wilderness areas: Neota, Rawah, Comanche Peak, and Cache la Poudre. Within each wilderness area, multiple patches are examined, with each patch having a uniform size. The dataset consists of 30,860 instances (i.e. patches) and 56 attributes, including an attribute representing the index of each instance. These attributes describe various environmental characteristics of the patches, such as elevation, slope, aspect, soil type, and wilderness area classification, which may be relevant to determining the forest cover type. A comprehensive data dictionary is provided in Appendix 3, detailing each attribute's description, data type (e.g., integer, float, string), and valid range. Additionally, some columns may contain invalid or inconsistent values, which will be addressed in the data cleaning section.Lastly, for convenience, I have also included two new columns from the cleaned dataset in the data dictionary and highlighted them in bold.

3.1 Data ingestion

To ingest this CSV dataset, I used Python's Pandas library and stored it as a Pandas DataFrame. Specifically, I employed the pandas.read_csv() function to load the dataset from "forest_cover.csv", ensuring that all data, including the index column, was properly stored. Since the dataset is contained within a single CSV file, only one DataFrame was created for this analysis. Pandas automatically detects the data type of each column based on its values. For instance, columns containing only integer values are assigned the int type, while those containing both integers and floats are assigned the float type. Certain columns exhibit different data types due to missing values (NaN); for example, "Soil_Type7" to "Soil_Type11" contain only integers and remain as int, whereas other "Soil_Type" columns contain NaN values, causing Pandas to interpret them as float. The dataset consists of 56 attributes (including the index column), which can be broadly categorized into interval, nominal, and ratio attributes. The interval attribute "Aspect" is stored as a float, while ratio attributes such as "Elevation," "Slope," "Horizontal_Distance_To_Hydrology," "Vertical_Distance_To_Hydrology," "Horizontal_Distance_To_Roadways," "Hillshade_9am," "Hillshade_Noon," "Hillshade_3pm," and "Horizontal_Distance_To_Fire_Points" are stored as either int or float, depending on the specific data. Nominal attributes include "Soil_Type1" through "Soil_Type40" and the wilderness area columns ("Neota," "Rawah," "Comanche Peak," and "Cache la Poudre"). These attributes follow a one-hot encoding scheme, where only one column in each group can have a value of 1 while all others remain 0. Due to the presence of NaN values, Pandas may store them as either int or float. Another nominal attribute, "Forest_Cover," is stored as an object type, as Pandas recognizes non-numeric columns as object by default.

3.2 Data quality assurance and cleaning

The dataset contains several data quality issues that need to be addressed. With the exception of the response variable "Forest Cover", the wilderness area columns ("Neota," "Rawah," "Comanche Peak," and "Cache la Poudre") and the columns "Soil_Type7" to "Soil_Type11", all other columns contain missing values. Additionally, the "Soil_Type" columns use one-hot encoding, which is inefficient and could be represented more effectively in a single column instead of 40 separate columns. The same issue applies to the "Neota", "Rawah", "Comanche Peak", and "Cache la Poudre" columns. Several columns, such as "Hillshade_9am", "Hillshade_Noon", "Hillshade_3pm", "Horizontal_Distance_To_Hydrology", "Horizontal_Distance_To_Roadways", "Horizontal_Distance_To_Fire_Points", and "Aspect", contain values that fall outside of valid ranges. For example, negative horizontal distances are present, which are not meaningful or valid in this context. Finally, approximately 80% of the data is focused on the “Spruce/Fir” and “Lodgepole Pine” forest cover types. This imbalance may introduce bias into the analysis, potentially limiting the generalizability of findings for other tree species.

To address the missing values in the "Elevation" column, I first analyzed the average elevation for different wilderness areas using a combination of dropna(), groupby(), and mean() functions, counting only rows with valid elevation data. I then filled the missing values by using the average elevation for the corresponding wilderness area. This approach makes sense because wilderness areas generally have minimal elevation differences, thus preserving the original trends in the data.

For the missing values in the "Soil_Type" columns, I first observed that only one column per row could have a value of 1, ensuring that each row belongs to a single soil type. In such cases, I filled the NaN with 0 when a row already had a soil type. For rows lacking a soil type, I categorized the data based on wilderness areas, as each wilderness area has a distinct soil type.I calculated the distribution of soil types in each wilderness area and then filled in the missing values based on the wilderness area type of the current row. This ensured that the data pattern remained intact, and I used the random.choices() function to assign soil types. This method is valid as most wilderness areas have a majority of valid rows. Once the missing values were filled, I combined the individual soil type columns into one "Soil_Type" column with an int datatype, which ranges from 1 to 40, for ease of analysis.

For missing values in the "Hillshade_9am”, "Hillshade_Noon”, and "Hillshade_3pm" columns, I filled in the missing values by using the mean of the other two columns if only one value was missing. If two values were missing, I used the remaining value to fill the others. For rows with all three missing values, I dropped those row, as they accounted for only 320 rows. This approach is reasonable because the hill shade values do not change much from 9 am to 3 pm. I also replaced illegal values in these columns, such as negative numbers and values greater than 255, by setting them to 0 and 255, respectively.

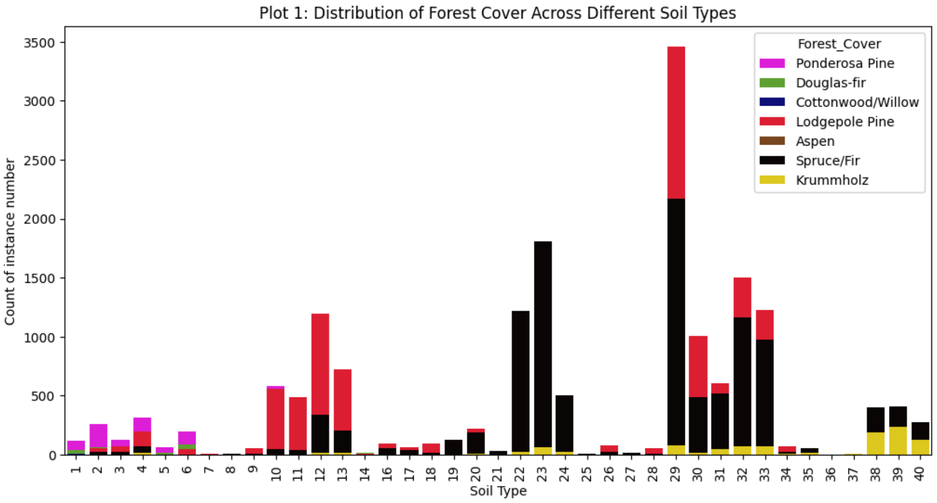
I dropped the "Vertical_Distance_To_Hydrology" column after analyzing its distribution, as it showed minimal variation. Regarding missing values in the "Horizontal/Vertical_Distance_To_XXXX” columns, they were filled with their respective means directly, as these columns did not seem strongly related to wilderness areas. I used the iterrows() function and loc[] method to fill missing values and change negative values to positive by applying the abs() function. This ensured that all horizontal distances were positive, preserving the logical consistency of the data.

For the "Aspect" and "Slope" columns, I decided to drop them as the mean values for each tree type were very similar, and their removal did not significantly affect the dataset. I used the drop() function to eliminate these columns.

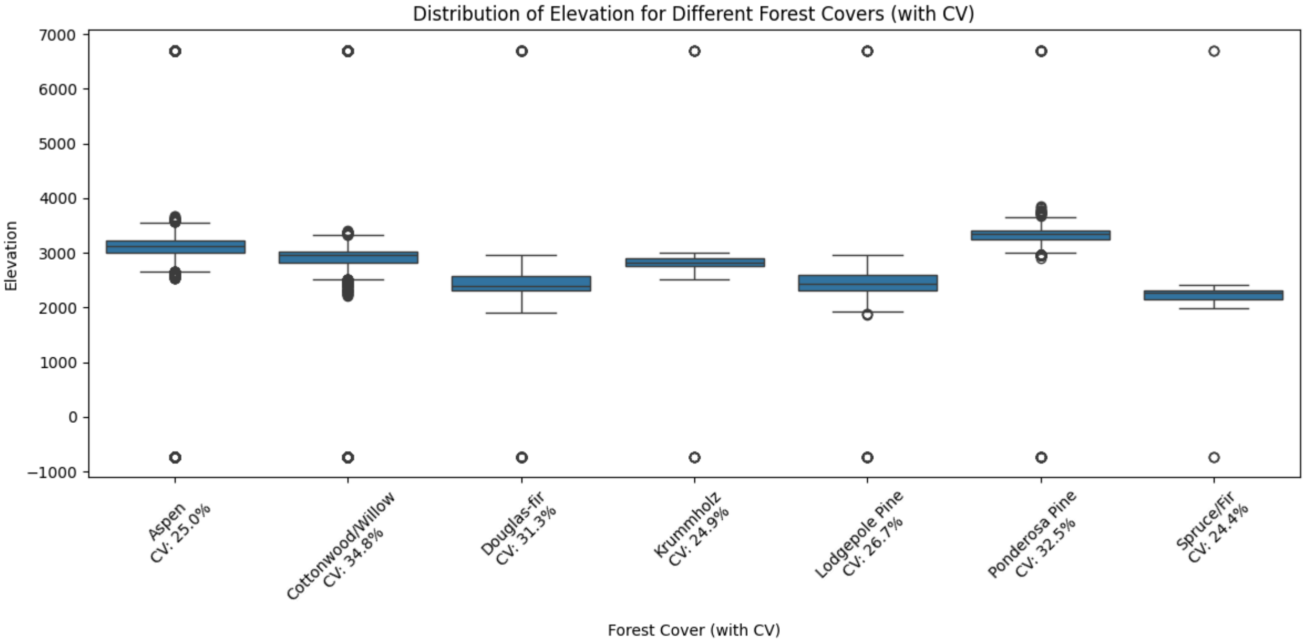
For the "Neota”, "Rawah”, "Comanche Peak”, and "Cache la Poudre" columns, which were one-hot encoded, I simply filled missing values with 0, as each row had exactly one of these columns with a value of 1. I also combined these columns into one column named "Wilderness_Area”, with a string datatype. The values in this column can only be one of "Neota”, "Rawah”, "Comanche Peak”, or "Cache la Poudre" .

In summary, I utilized a combination of analytical techniques, including grouping, probability distributions, and random selection, to handle missing values and correct invalid entries. Additionally, I applied various functions such as dropna(), groupby(), apply(), and iterrows() to clean the data while preserving its integrity for further analysis. Lastly, I retained columns that were not directly relevant to my research question to maintain as much data as possible for Assignment 2.

4. Exploratory Data Analysis


Plot 1: Distribution of Forest Cover Across Different Soil Types

To investigate the research question, I first explore the relationship between forest cover (tree species) and soil type. In Plot 1, I used stacked bar plots to illustrate the distribution of forest covers across different soil types. This method was chosen because both variables are nominal, and it is important to visualize the count of each type. The y-axis represents the count, and the stacked bars effectively show how tree species are distributed across soil types and wilderness areas. From this plot, we can conclude that tree species do show a strong relationship with soil types. For example, Lodgepole Pine tends to prefer soil types 10, 11, 12, 13, 29, and 30, while Spruce/Fir is more commonly found in soil types 19, 20, 22, 23, 24, 29, 31, 32, and 33. Ponderosa Pine prefers soil types 1, 2, 4, and 6, and Krummholz is typically found in soil types 38, 39, and 40. These patterns indicate that certain tree species are indeed more suited to specific soil types. However, for other tree species with limited data, we cannot draw clear conclusions from the plot. It's also important to note that the majority of the data is focused on Lodgepole Pine and Spruce/Fir, so these conclusions are particularly reliable for these two species, while the results for other species may not be as accurate.



Now, let’s use Plot 2 to explore the relationship between elevation and different tree species. In this plot, I used a box plot to represent the relationship between a ratio variable (elevation) and a nominal variable (forest cover type). The X-axis represents the forest cover type (tree species) with its Coefficient of Variation (CV), and the Y-axis represents elevation. From the plot, we can observe that most tree species are not concentrated around a specific elevation, as indicated by the wide distance between Q1 and Q3 for each species. The CV further supports this, with all CV values greater than 20%, indicating that the elevation data is spread out and not concentrated around the mean. However, we do see that most tree species tend to be clustered between elevations of 2200 and 3600 meters, suggesting that while there isn't a strong correlation, elevation may still play a role in tree distribution. In conclusion, while there is some relationship between tree species and elevation, it is not strong enough to make definitive claims, and elevation alone does not strongly determine forest cover type.

In summary, there is a strong relationship between soil type and forest cover type, as demonstrated in Plot 1, where specific tree species are clearly associated with certain soil types. However, the relationship between forest cover and elevation in Plot 2 is less apparent. Trees can thrive across a wide range of elevations, and some outliers even exceed this range, indicating that elevation does not have a strong influence on forest cover type.

It’s important to note that due to the imbalance in the dataset, with Lodgepole Pine and Spruce/Fir dominating the distribution, the conclusions drawn from this analysis may be biased. Since the majority of the data is focused on these two tree species, the findings regarding other tree types may not be as reliable. This bias could affect subsequent analysis or model-building tasks. For instance, the distribution of soil types and elevations may not be representative for less frequent tree species, potentially leading to overgeneralized conclusions. To address this imbalance in the machine learning model for Assignment 2, one potential solution is to assign higher weights to the underrepresented tree species during model training. However, even with this approach, certain tree species that prefer specific soil types, which are not well represented in the dataset, may still not be adequately covered. While weighting can help mitigate some of the imbalance, it may not fully resolve the issue if the data for particular soil types or tree species remains sparse.

Group Discussion

Among the three datasets, the auto_price dataset exhibits the highest proportion of missing values, affecting both numerical and categorical columns. Key features such as Price (17.99%), Fuel (38.99%), and Mileage (8.99%) show considerable missing rates, while others like Extras and Upholstery_Type have over 80% missing values. Despite these gaps, the dataset is well-structured, with 18,286 rows and 32 columns, and includes all relevant variables for its research objective, particularly Price_Value (the target variable), Mileage_Value, Age, Fuel, and Body_Type, making it highly suitable for predicting vehicle prices. However, the dataset's high degree of skewness in numerical variables, such as Price and Mileage, and its imbalanced categorical features—where sedans and SUVs are overrepresented while convertibles and electric vehicles are underrepresented—may introduce bias into research results. This imbalance could lead to favoring more common vehicle types, potentially limiting insights into price determinants for less common vehicle types.

The diabetes dataset is the largest, containing approximately 264,802 rows and 22 columns, with each row representing an individual's health record. While its large size allows for training complex models, missing values in several columns—such as PhysActivity, Fruits, and Veggies (with over 70% missing data in some)—affect its completeness. However, it contains medically significant features, including Age, BMI, DiffWalk, and Diabetes status, which are essential for health prediction tasks. These attributes provide valuable insights into diabetes risk factors and patient health conditions. Moreover, the dataset's relatively balanced distribution (65% non-diabetic and 35% diabetic) supports reliable binary classification and generalization, making it well-suited for health-related research.

The forest_cover dataset, in contrast, has relatively few missing values. While some Soil_Type variables exhibit missing rates between 20%-60%, the dataset remains mostly complete for its primary features. It comprises 30,860 rows and 56 columns, with each row representing a patch within a wilderness area. Notably, only the forest_cover dataset's response attribute remains fully intact, with no missing values, unlike the other two datasets where the response variables are affected by missing data. This makes the forest_cover dataset particularly reliable for research. However, its scope is limited by the absence of key ecological parameters, such as rainfall and temperature, which significantly narrows the range of possible research questions. Additionally, the dataset is highly imbalanced, with dominant species such as Lodgepole Pine and Spruce/Fir accounting for around 85% of the data, while rare species like Cottonwood/Willow make up less than 1%. This severe imbalance could cause research results to be biased toward the dominant species, making it difficult to accurately predict rare species or generalize findings to less common habitats. This imbalance may distort the ecological insights derived from the data.

In conclusion, the auto_price dataset offers a well-structured set of variables for predicting vehicle prices, but its high proportion of missing values and class imbalance may introduce biases, particularly for less common vehicle types. The diabetes dataset, with its large size and balanced class distribution, is ideal for health-related research, supporting reliable insights into diabetes risk factors and patient health conditions. On the other hand, the forest_cover dataset, although relatively clean and manageable in size, is limited by the absence of key ecological variables and a significant class imbalance. These factors make it challenging to fully explore the influences on forest cover distribution and ecological dynamics.

Building on the comparison of the three datasets, each required specific exploratory data analysis (EDA) techniques tailored to their unique characteristics and research objectives.

The auto_price dataset, focused on car price prediction, used scatter plots and boxplots to explore the relationships between continuous variables, such as mileage and age, and the target variable, price. These methods proved effective in visualizing trends and detecting outliers, which are essential for regression modeling. Scatterplots were particularly useful in capturing linear relationships between two ratio variables. For instance, the analysis showed that price is negatively correlated with mileage. However, when trying to represent too many categories, such as the working years of a car, the scatter plot became cluttered, making it difficult for the reader to distinguish between the different color-coded dots. Boxplots, on the other hand, provided valuable insights into the distribution and variance of the target variable, helping to determine whether the data tends to concentrate or spread out. For example, the boxplot revealed that coupe prices are more concentrated compared to vans. However, scatterplots may fail to capture non-linear relationships, and boxplots can become less informative when dealing with variables with numerous outliers or skewed distributions, potentially overlooking subtle patterns. Specifically, boxplots do not give an exact distribution of the data, such as whether the data follows a normal distribution or if there are other local peaks.

The diabetes dataset, aimed at predicting diabetes risk, used bar charts and heatmaps to explore the relationships between lifestyle factors, such as BMI and age, and the risk of diabetes. Bar charts were helpful in identifying patterns for categorical variables. For example, we used bar charts to demonstrate that age group and diabetes rate are positively related. Heatmaps, on the other hand, clearly revealed relationships between two nominal attributes with multiple values. For instance, we used the heatmap to identify a strong correlation between poor general health and obesity. A common issue with bar charts is that they become unclear when there are too many categories for a nominal variable, making it hard to detect differences between each bar. Additionally, bar charts cannot predict multiple attributes simultaneously. For this research, the second issue was particularly relevant—while we could identify trends, the steepness of the change wasn't always clearly illustrated. We discussed adding a line to indicate the trend more clearly. From our prior knowledge, we know that heatmaps tend to underperform with smaller datasets. However, this wasn't a problem in this case, as the dataset was sufficiently large. One limitation of heatmaps is that they do not show the detailed relationship between variables, such as whether the relationship is positive or negative.

For the forest_cover dataset, which focuses on forest cover distribution, stacked bar charts and boxplots were used to examine the relationship between tree species, soil types, and elevation. The stacked bar chart was useful for exploring the distribution of tree species across different soil types (two nominal variables related by the count of instances). This allowed the reader to easily identify which trees are more likely to thrive in specific soil types. A common issue with stacked bar charts is that when there are many categories within a nominal variable, it can become difficult to distinguish between them due to the different colors representing each category. However, this wasn't a problem in this dataset, as there are only two dominant tree species. We also carefully selected distinct colors, like black and red, to ensure they were easily distinguishable. As for boxplots, while they helped determine whether data tends to concentrate or spread out, they didn't allow for a precise understanding of the exact distribution of the data.

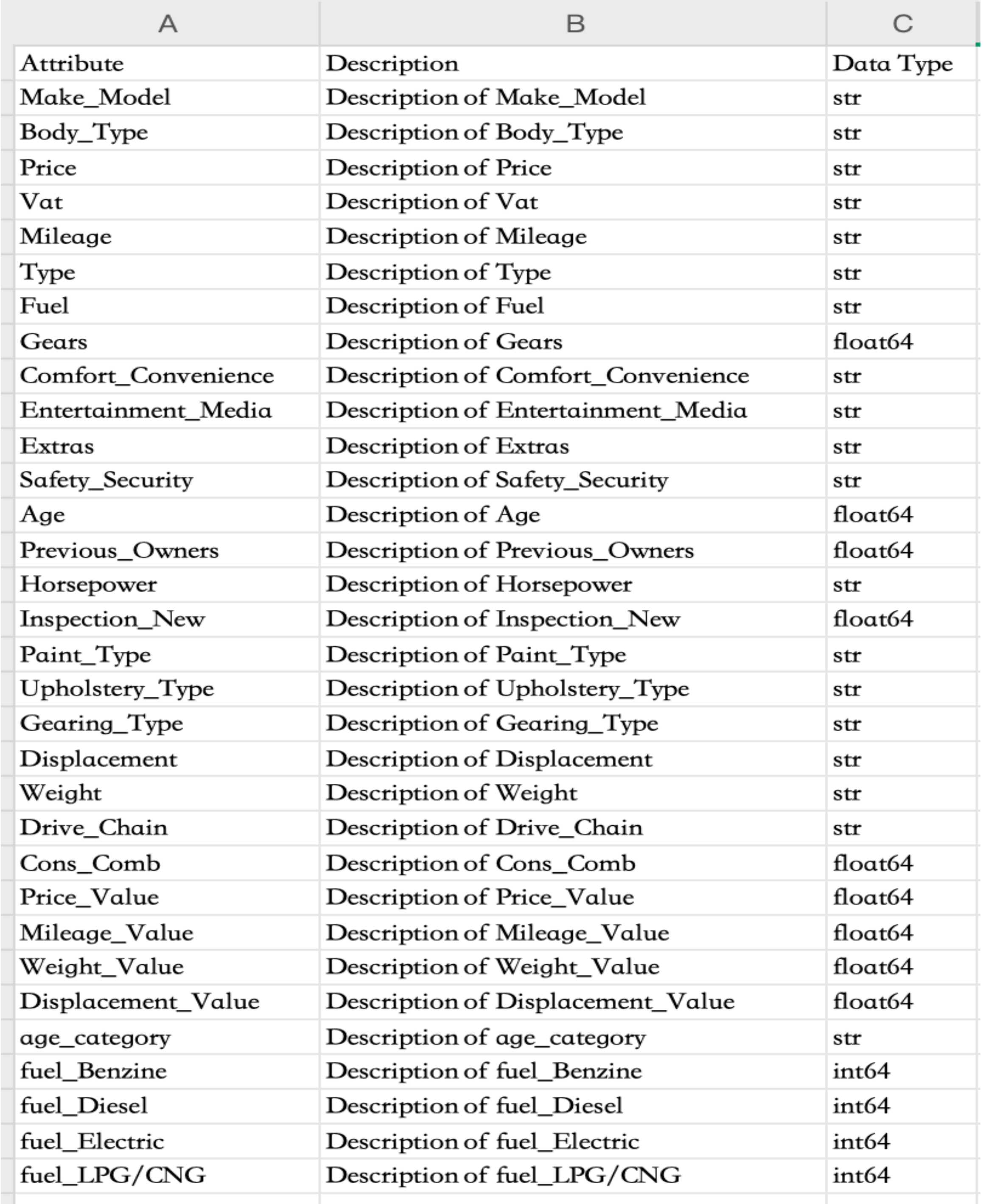
Group Conclusion

The exploratory data analysis (EDA) conducted by all group members provided valuable insights into the structure and key patterns within each dataset. For the auto_price dataset, we identified strong relationships between vehicle characteristics and price, with mileage showing a clear negative correlation and engine displacement a positive one. The diabetes dataset revealed that non-glucose features, such as muscle fatigue, ankle reflex, and cholesterol levels, play an important role in predicting diabetes, highlighting the potential for broader health assessments beyond traditional glucose measurements. The forest_cover dataset demonstrated a strong link between soil type and tree species distribution, while elevation showed no significant association.

These findings directly contribute to answering our research questions by identifying the most relevant variables for prediction and analysis. Understanding key factors influencing vehicle prices supports accurate price estimation, while the identified health indicators in the diabetes dataset offer valuable insights for developing non-invasive screening models. Similarly, the forest_cover dataset provides useful ecological insights for modeling species distribution based on soil properties.

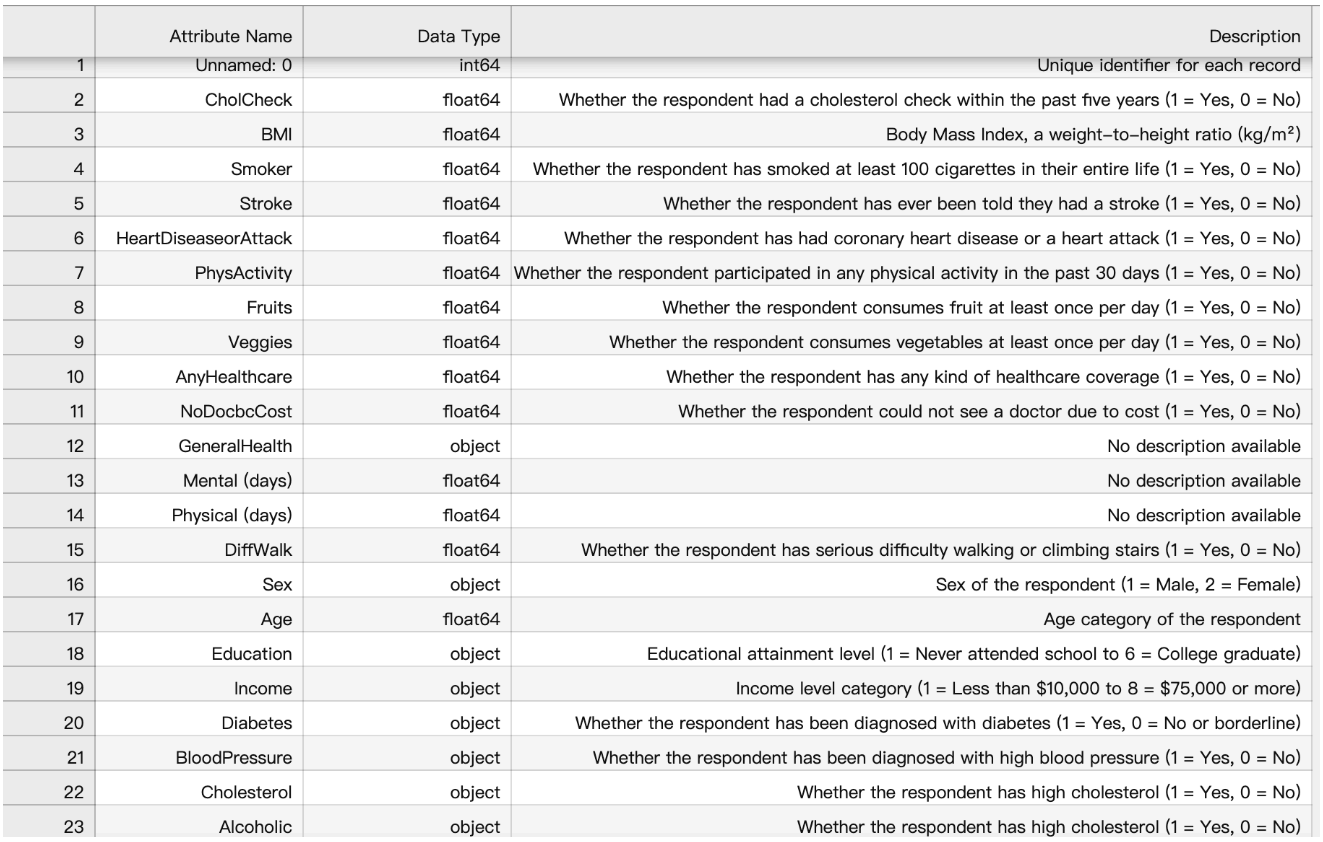
Based on group discussions, we have decided to proceed with the diabetes dataset for the next stage of our project. This decision was made for two main reasons. First, the diabetes dataset offers the largest number of instances, providing a stronger foundation for statistical analysis and model training. Its more balanced distribution also reduces the risk of bias, ensuring the development of a fair and reliable predictive model. Second, the EDA results played a significant role. The forest_cover dataset presented weaknesses in the EDA, particularly in analyzing rare tree species due to data skewness. Similarly, the auto_price dataset showed that most data is skewed toward cars with low mileage, which makes the conclusions less representative. Therefore, we have ultimately chosen the diabetes dataset for further analysis.

Appendix 1



A	B	C
Attribute	Description	Data Type
Make_Model	Description of Make_Model	str
Body_Type	Description of Body_Type	str
Price	Description of Price	str
Vat	Description of Vat	str
Mileage	Description of Mileage	str
Type	Description of Type	str
Fuel	Description of Fuel	str
Gears	Description of Gears	float64
Comfort_Convenience	Description of Comfort_Convenience	str
Entertainment_Media	Description of Entertainment_Media	str
Extras	Description of Extras	str
Safety_Security	Description of Safety_Security	str
Age	Description of Age	float64
Previous_Owners	Description of Previous_Owners	float64
Horsepower	Description of Horsepower	str
Inspection_New	Description of Inspection_New	float64
Paint_Type	Description of Paint_Type	str
Upholstery_Type	Description of Upholstery_Type	str
Gearing_Type	Description of Gearing_Type	str
Displacement	Description of Displacement	str
Weight	Description of Weight	str
Drive_Chain	Description of Drive_Chain	str
Cons_Comb	Description of Cons_Comb	float64
Price_Value	Description of Price_Value	float64
Mileage_Value	Description of Mileage_Value	float64
Weight_Value	Description of Weight_Value	float64
Displacement_Value	Description of Displacement_Value	float64
age_category	Description of age_category	str
fuel_Benzine	Description of fuel_Benzine	int64
fuel_Diesel	Description of fuel_Diesel	int64
fuel_Electric	Description of fuel_Electric	int64
fuel_LPG/CNG	Description of fuel_LPG/CNG	int64

Appendix 2



	Attribute Name	Data Type	Description
1	Unnamed: 0	int64	Unique identifier for each record
2	CholCheck	float64	Whether the respondent had a cholesterol check within the past five years (1 = Yes, 0 = No)
3	BMI	float64	Body Mass Index, a weight-to-height ratio (kg/m²)
4	Smoker	float64	Whether the respondent has smoked at least 100 cigarettes in their entire life (1 = Yes, 0 = No)
5	Stroke	float64	Whether the respondent has ever been told they had a stroke (1 = Yes, 0 = No)
6	HeartDiseaseorAttack	float64	Whether the respondent has had coronary heart disease or a heart attack (1 = Yes, 0 = No)
7	PhysActivity	float64	Whether the respondent participated in any physical activity in the past 30 days (1 = Yes, 0 = No)
8	Fruits	float64	Whether the respondent consumes fruit at least once per day (1 = Yes, 0 = No)
9	Veggies	float64	Whether the respondent consumes vegetables at least once per day (1 = Yes, 0 = No)
10	AnyHealthcare	float64	Whether the respondent has any kind of healthcare coverage (1 = Yes, 0 = No)
11	NoDocbcCost	float64	Whether the respondent could not see a doctor due to cost (1 = Yes, 0 = No)
12	GeneralHealth	object	No description available
13	Mental (days)	float64	No description available
14	Physical (days)	float64	No description available
15	DiffWalk	float64	Whether the respondent has serious difficulty walking or climbing stairs (1 = Yes, 0 = No)
16	Sex	object	Sex of the respondent (1 = Male, 2 = Female)
17	Age	float64	Age category of the respondent
18	Education	object	Educational attainment level (1 = Never attended school to 6 = College graduate)
19	Income	object	Income level category (1 = Less than $10,000 to 8 = $75,000 or more)
20	Diabetes	object	Whether the respondent has been diagnosed with diabetes (1 = Yes, 0 = No or borderline)
21	BloodPressure	object	Whether the respondent has been diagnosed with high blood pressure (1 = Yes, 0 = No)
22	Cholesterol	object	Whether the respondent has high cholesterol (1 = Yes, 0 = No)
23	Alcoholic	object	Whether the respondent has high cholesterol (1 = Yes, 0 = No)

Appendix 3

Attribute Name	Description	Data Type	Range
Elevation	Elevation of current patch in meters.	float64	any float number
Aspect	Describes the orientation of a slope relative to cardinal directions.	float64	[0:360]
Slope	The steepness or incline of the patch, measured in degrees.	float64	any float number
Horizontal_Distance_To_Hydrology	Horizontal distance to nearest surface water features, measured in meters.	float64	non-negative float number
Vertical_Distance_To_Hydrology	Vertical distance to nearest surface water features, measured in meters.	float64	any float number
Horizontal_Distance_To_Roadways	Horizontal distance to the nearest roadway, measured in meters.	float64	non-negative float number
Horizontal_Distance_To_Fire_Points	Horizontal distance to nearest wildfire ignition points, measured in meters.	float64	non-negative float number
Hillshade_9am	Hill shade index at 9 am, summer solstice. 0 means complete darkness, 255 means full sunlight.	float64	[0:255]
Hillshade_Noon	Hill shade index at noon, summer solstice. 0 means complete darkness,	float64	[0:255]

	255 means full sunlight.		
Hillshade_3pm	Hill shade index at 3 pm, summer solstice. 0 means complete darkness, 255 means full sunlight.	float64	[0:255]
Soil_Type1 & … & Soil_Type40	These 40 columns represent the soil type of the current patch. Each patch can only have one type of soil, meaning only one of the Soil_Type columns can be set to 1 at a time, with all other columns set to 0. A value of 1 indicates the presence of a specific soil type, while a value of 0 indicates its absence.	float64	either 0 or 1
Forest_Cover	The response variable, representing the type of forest cover in the patch.	object	It should be one of the following: 'Spruce/Fir', 'Lodgepole Pine', 'Ponderosa Pine', 'Aspen', 'Douglas-fir', 'Krummholz', or 'Cottonwood/Willow'.
Neota	Indicates whether the patch is located in a specific wilderness area, with 1 indicating belonging and 0 indicating not belonging. Each patch can belong to only one wilderness area.	int64	either 0 or 1.

Rawah	Indicates whether the patch is located in a specific wilderness area, with 1 indicating belonging and 0 indicating not belonging. Each patch can belong to only one wilderness area.	int64	either 0 or 1.
Comanche Peak	Indicates whether the patch is located in a specific wilderness area, with 1 indicating belonging and 0 indicating not belonging. Each patch can belong to only one wilderness area.	int64	either 0 or 1.
Cache la Poudre	Indicates whether the patch is located in a specific wilderness area, with 1 indicating belonging and 0 indicating not belonging. Each patch can belong to only one wilderness area.	int64	either 0 or 1.
Soil_Type	This is **a new column in the clean datase**t. It means the soil type of the current patch. For example 7 means soil type 7	int64	[1:40]
Wilderness_Area	This is **a new column in the clean datase**t. It means the corresponding wilderness area of the current patch.	object	It should be one of the following: 'Rawah', 'Comanche Peak', Cache la Poudre', or 'Neota''.