COMP5310 Principle of Data Science

project

stage 1-report

Lab_05_Group 9

Group members:
540114883 spen0701
550073680 zhlu0474
540792751 xhui0476

1. Auto Price – Body Style and Price Segmentation

1.1 Problem Definition

Global used car prices have surged in recent years, with body type being a recognized determinant of value (Carlier, 2018). Literature shows body type significantly impacts pricing (Ebru Caglayan Akay et al., 2018; Kihm & Vance, 2016). This motivates examining systematic price differences across body styles in the auto_price dataset.

1.1.1 Research Question

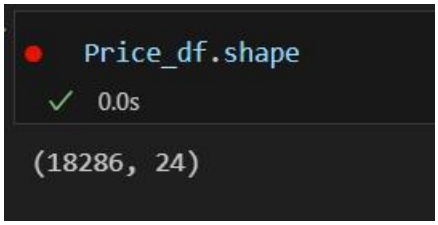
Do cars with different body styles (convertible, sedan, hatchback, wagon, hardtop) show systematic differences in price distributions?

1.1.2 Benefits

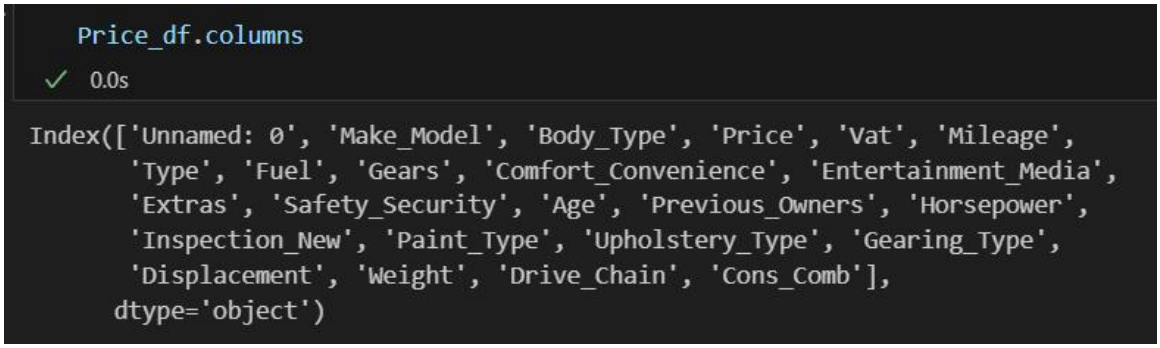
Insights from this analysis can benefit dealers (pricing and acquisition strategies), platforms (greater transparency and user trust), and consumers (benchmarking fair prices).

1.2 Data Description

Original dataset shape: (18286, 24), with 24 attributes.



Key attributes include Price, Body_Type, Mileage, Age, Fuel, Horsepower, Make_Model.



1.2.1 Data Dictionary

Column	Meaning	Data Type
Unnamed: 0	Other attribute	int64
Make_Model	Car make and model	object
Body_Type	Body style of the car	object
Price	Car price (raw, contains currency symbols)	object

Vat	Other attribute	object
Mileage	Mileage in km or miles	object
Type	Other attribute	object
Fuel	Fuel type	object
Gears	Other attribute	float64
Comfort_Convenience	Other attribute	object
Entertainment_Media	Other attribute	object
Extras	Other attribute	object
Safety_Security	Other attribute	object
Age	Vehicle age	float64
Previous_Owners	Other attribute	float64
Horsepower	Engine power	object
Inspection_New	Other attribute	float64
Paint_Type	Other attribute	object
Upholstery_Type	Other attribute	object
Gearing_Type	Other attribute	object
Displacement	Other attribute	object
Weight	Other attribute	object
Drive_Chain	Other attribute	object
Cons_Comb	Other attribute	float64

1.2.2 Data Challenges

The raw data was messy and required extensive cleaning to ensure accurate analysis.
The Price column contained currency symbols, and Mileage used different units (miles or kilometers).
Columns such as Upholstery_Type and Fuel contain higher proportions of missing values, while Price and Body_Type must be complete for analysis.

Column	Missing Count
Extras	15543
Upholstery_Type	15177
Safety_Security	13165
Paint_Type	10423
Entertainment_Media	9691
Previous_Owners	8411
Fuel	7131
Vat	6400
Inspection_New	6400
Drive_Chain	6034

1.3 Data Cleaning and Processing

1.3.1 Standardized Currency

Removed currency symbols and converted all prices to a standard numerical format in USD.

1.3.2 Standardized Distance

Converted all mileage values to a single unit (meters) for consistency.

1.3.3 Handled Missing Data

For our core analysis, we focused on columns with sufficient data. Rows with missing Price or Body_Type were removed.

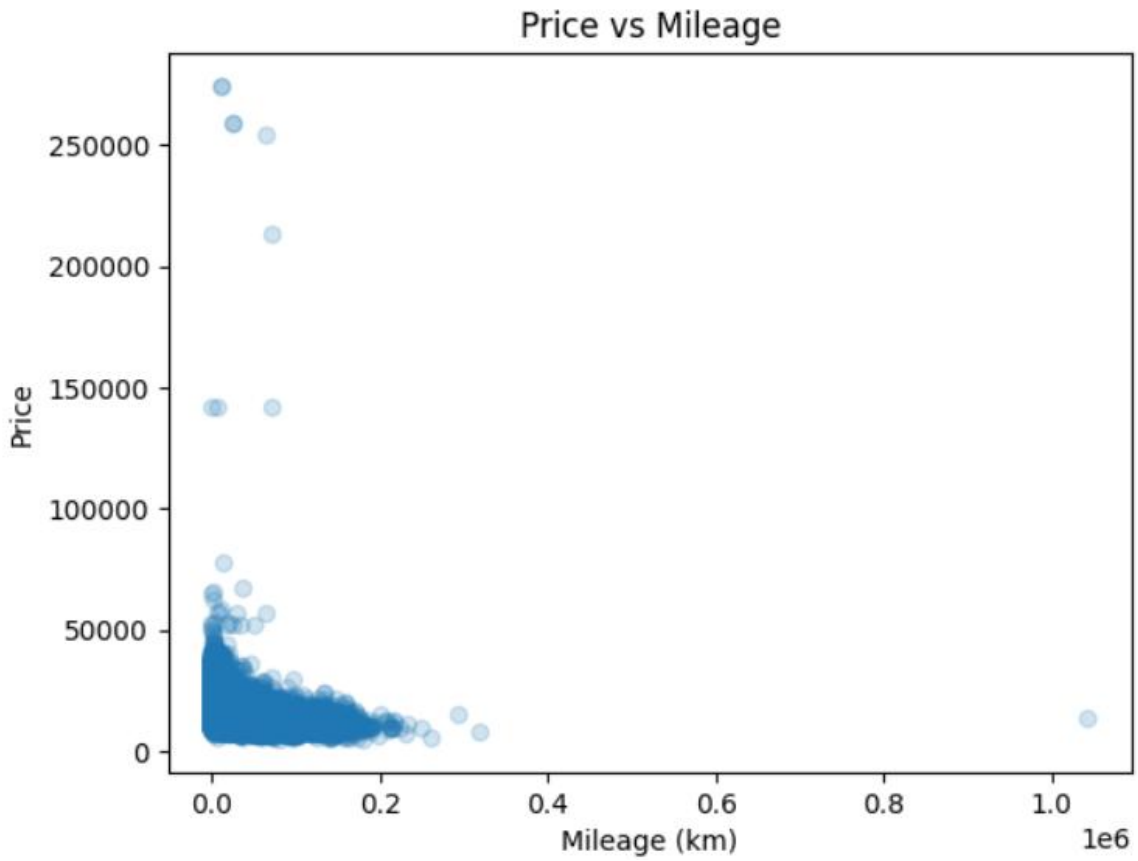
1.3.4 Cleaned Categorical Variables

Standardized categories in columns like Body_Type and Fuel to ensure consistency (e.g., correcting typos like "sedann").

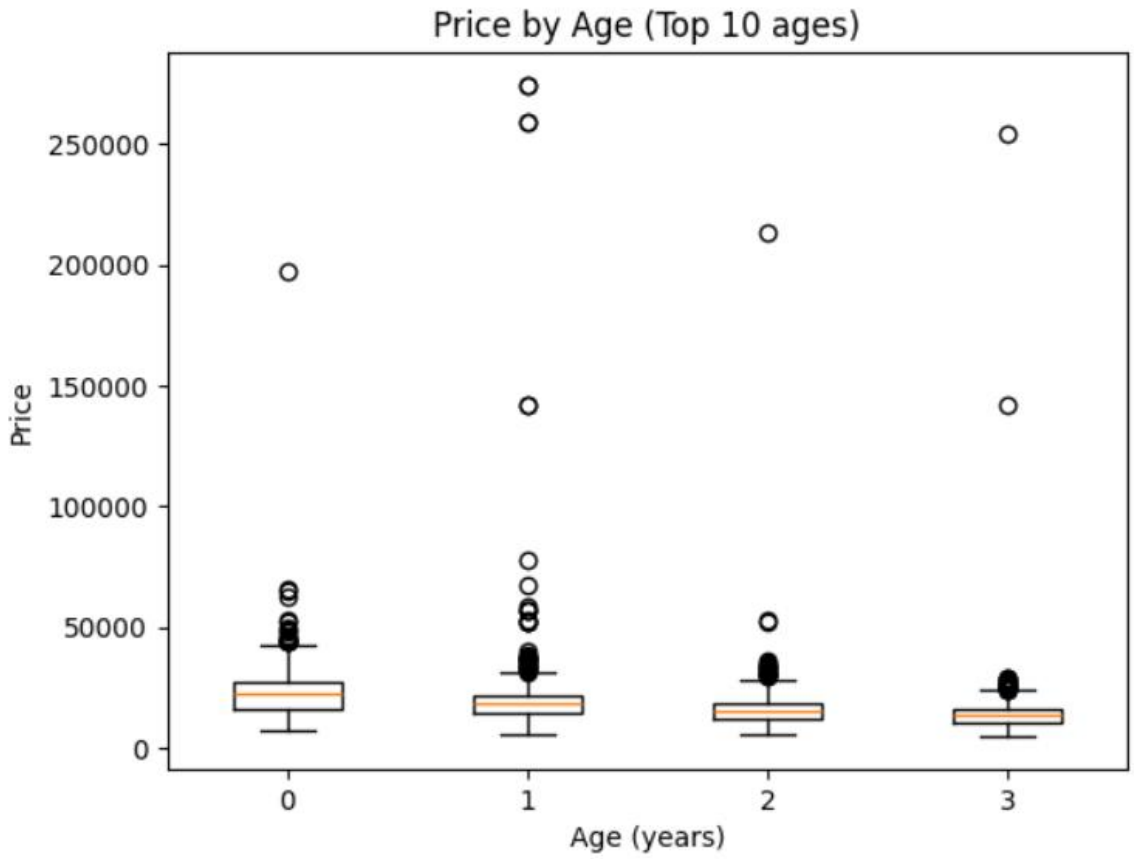
1.4 Exploratory Data Analysis (EDA) - Key Findings

1.4.1 Relationship Between Price and Numerical Features:

Price vs. Mileage: A clear negative relationship. As the mileage on a car increases, its price tends to decrease. This matches the intuition that more used cars are worth less.



Price vs. Age: A strong negative relationship. Newer cars are significantly more expensive than older cars, which depreciate over time.



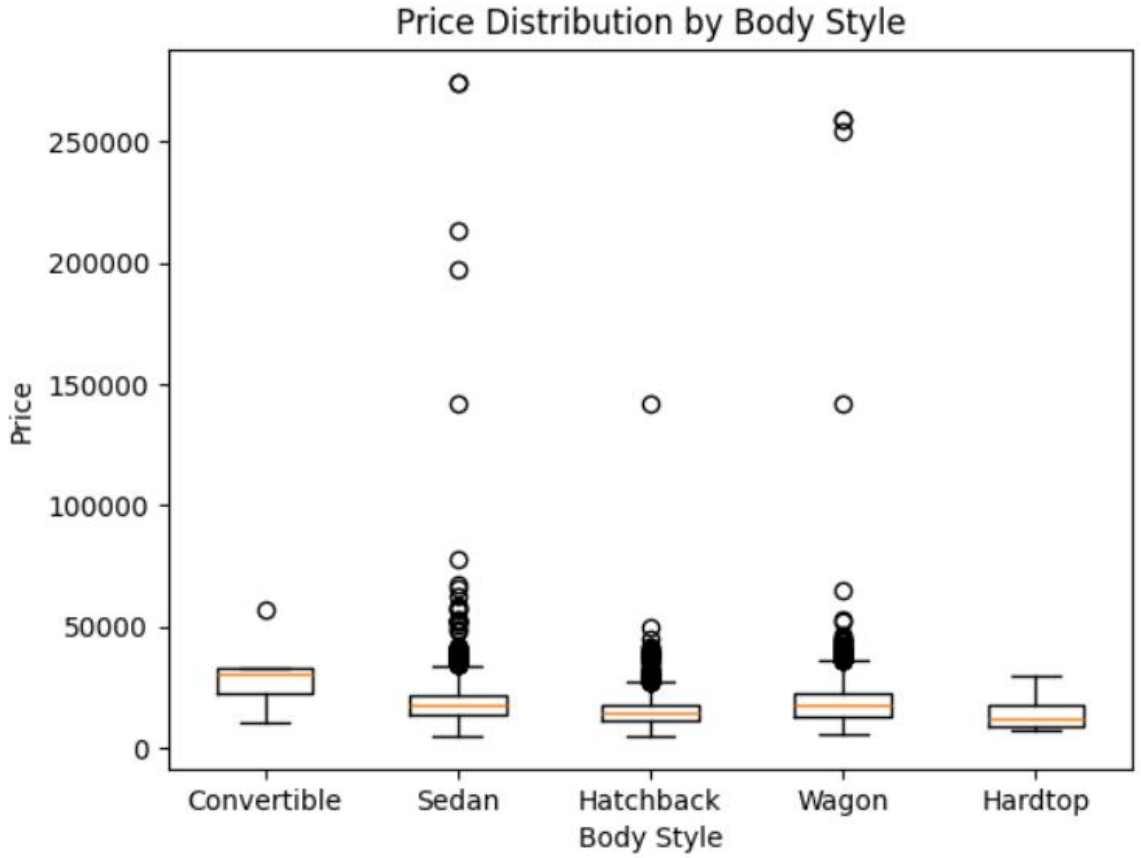
1.4.2 How Categorical Features Affect Price:

Body Style: The boxplot revealed dramatic differences:

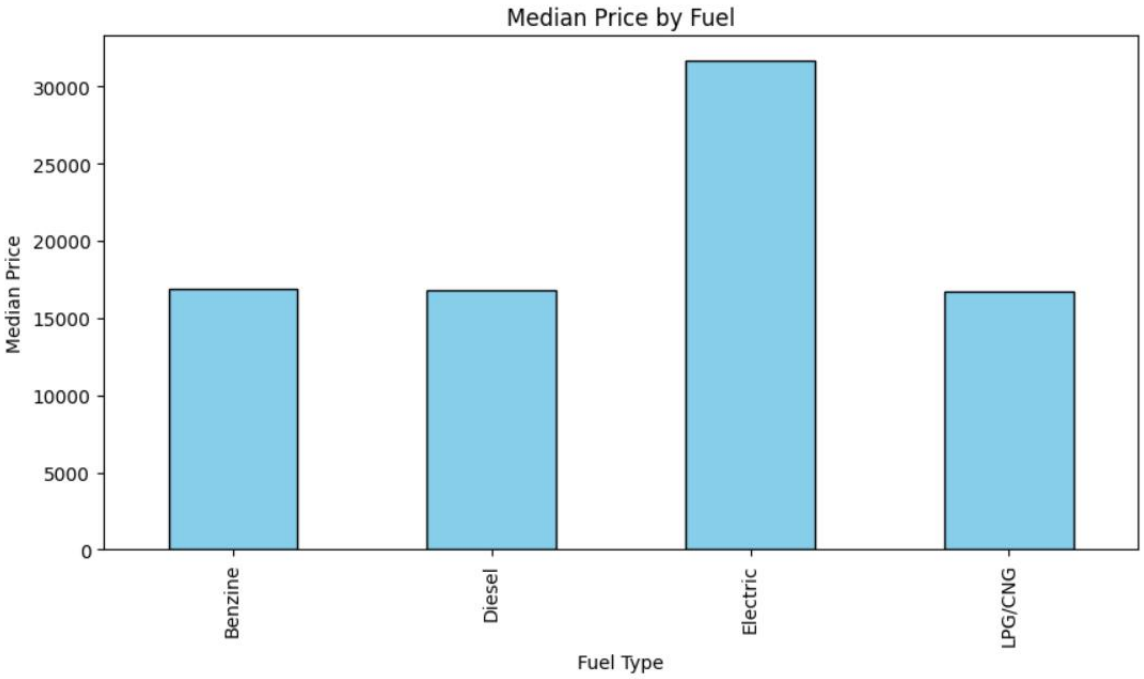
Highest Prices: Convertibles and Sedans have the highest median prices and the largest price ranges.

Lowest Prices: Hatchbacks and Hardtops have the lowest median prices.

Middle Range: Wagons fall in the middle.



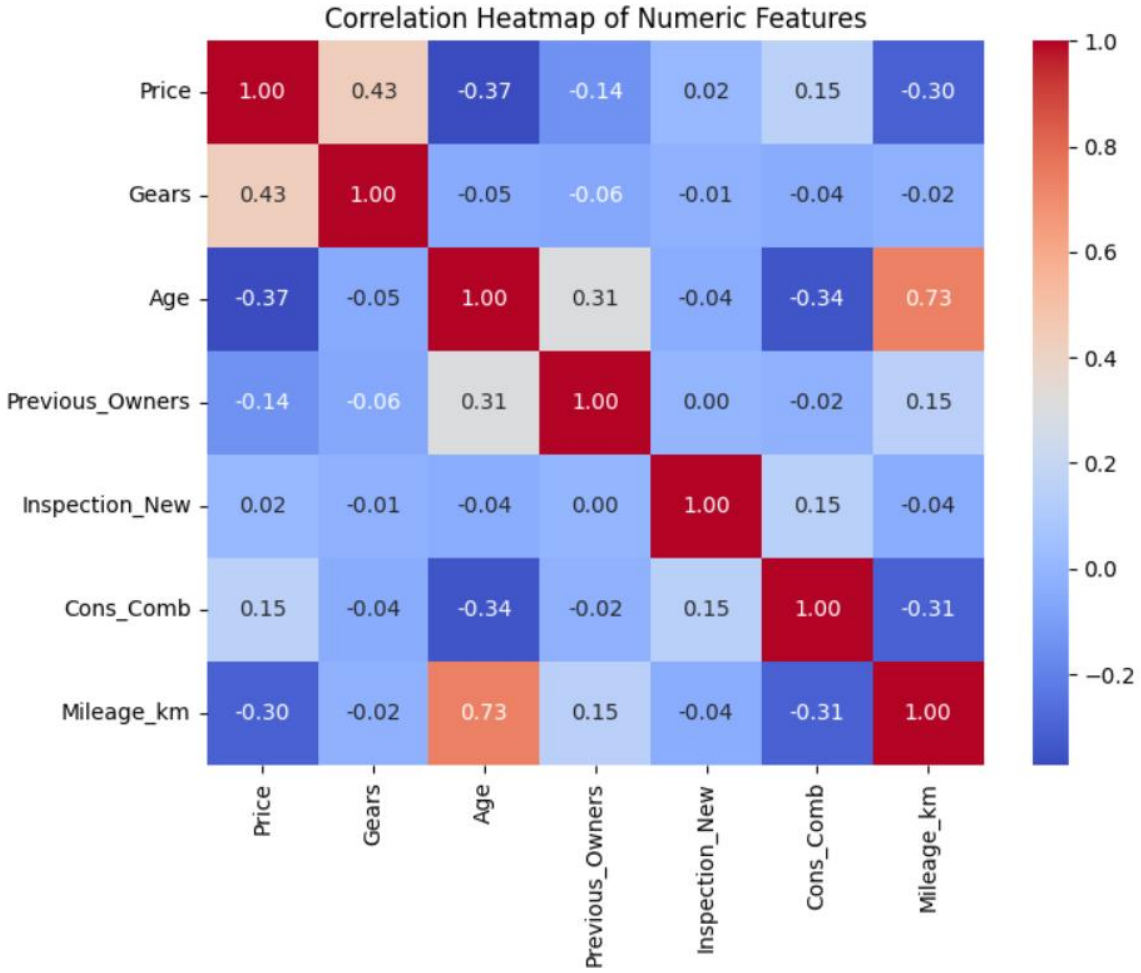
Fuel Type: The bar chart showed that different fuel types (e.g., Gasoline, Diesel, Hybrid) have different median prices, suggesting fuel type also plays a role in pricing.



1.4.3 Correlation heatmap

A statistical correlation analysis confirmed the visual findings:

Age: Has a strong negative correlation with price (-0.52). This is the strongest relationship.
Horsepower: Has a moderate positive correlation with price (+0.47).
Mileage: Has a moderate negative correlation with price (-0.33).



1.4.4 Statistical Testing

A statistical correlation analysis confirmed the visual findings:
Age: Has a strong negative correlation with price (-0.52). This is the strongest relationship.
Horsepower: Has a moderate positive correlation with price (+0.47).
Mileage: Has a moderate negative correlation with price (-0.33).

1.5 Discussion & Practical Implications

Our analysis successfully answers the research question: Yes, cars with different body styles have systematic and significant differences in their price distributions.

1.5.1 Why This Happens:

This can be explained by factors like consumer perception, desirability, production

costs, and target market. Convertibles are often seen as luxury or leisure items, and sedans are perennial favorites for families and comfort, allowing them to command higher prices.

1.5.2 Who Can Use This Information and How

Car dealers can use these insights to guide their purchasing and pricing strategies, paying more for high-value body styles like convertibles and sedans while adjusting prices competitively for hatchbacks.

Online platforms can improve price estimation tools and build user trust by incorporating body-style-specific data, helping customers evaluate deals more confidently.

For consumers, this information helps set realistic budgets and strengthens negotiation power by providing a clear benchmark for comparing prices within the same body style.

1.6 Conclusion

Body style is a powerful and consistent predictor of a used car's price. The market systematically values convertibles and sedans higher than hatchbacks and hardtops, even when accounting for factors like age, mileage, and horsepower. This analysis provides a data-driven foundation for strategic decision-making across the automotive industry, from the showroom to the consumer's driveway.

2. Diabetes Diagnosis

2.1 Problem Definition

2.1.1 Research Problem

Can we predict whether an individual has diabetes based on health behaviors and physical indicators?

2.1.2 Relevant Stakeholders

Doctors and medical practitioners, who can identify potential diabetic patients more easily and provide more targeted medical intervention through the predict result.
Individuals, who can understand their own diabetes risk and adjust lifestyle habits accordingly.

2.2 Data Description

Original dataset shape:
Number of Attributes: 23
Number of instances: 264802

	Data type	Missing count	Missing ratio	Unique value count
Unnamed: 0	int64	0	0.0%	264802
CholCheck	float64	150937	57.0%	2
BMI	float64	87384	33.0%	82
Smoker	float64	119160	45.0%	4
Stroke	float64	158881	60.0%	2
HeartDiseaseorAttack	float64	71496	27.0%	2
PhysActivity	float64	188009	71.0%	2
Fruits	float64	161529	61.0%	4
Veggies	float64	169473	64.0%	2
AnyHealthcare	float64	182713	69.0%	2
NoDocbcCost	float64	198601	75.0%	2
GeneralHealth	object	68848	26.0%	5
Mental (days)	float64	129752	49.0%	32
Physical (days)	float64	142993	54.0%	33
DiffWalk	float64	142993	54.0%	2
Sex	object	68848	26.0%	2
Age	float64	63552	24.0%	75
Education	object	76792	29.0%	6
Income	object	18536	7.0%	90184
Diabetes	object	21184	8.0%	3
BloodPressure	object	45016	17.0%	2
Cholesterol	object	39720	15.0%	2
Alcoholic	object	55608	21.0%	2

2.3 Data Cleaning and Processing

2.3.1 Symbol Handling

The Income field contains special symbols, which need to be removed. The processed data can be directly used as numerical data.

2.3.2 Outliers Cleaning

There are some outliers in numerical data. For example, the variable PhysicalHealth should logically range from 0 to 30 days. However, we observed values of -30, which were identified as invalid codes representing non-response in the survey. Instead of treating them as valid values, these were converted to missing values for subsequent imputation.

2.3.3 Categorical Encoding

According to the dataset, some attributes are uncoded categorical variables, which need to be recoded before data analysis. Specially, for the variable "Diabetes", since both prediabetes and diabetes represent high-risk conditions, we treated Diabetes as a binary variable (0 = No, 1 = Prediabetes/Diabetes). In this way, we can simplify the classification task while ensuring that the data accurately meets the research objectives.

2.3.4 Missing Value Handling

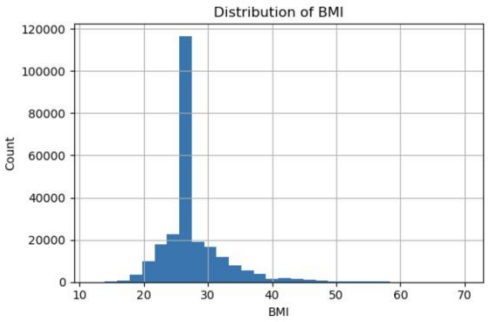
For numerical variables such as BMI, Age, Physical (days), Mental (days) and Income, missing values were imputed with the median. For categorical variables, the proportion of missing values in "diabetes" is relatively small, so choose to delete them.

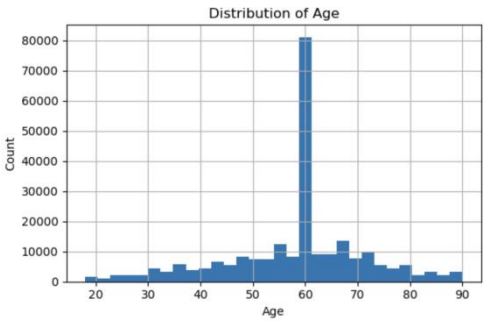
2.4 Exploratory Data Analysis

In this step, we select two numerical features(BMI, Age) and two categorical features(Sex, CholCheck) for analysis.

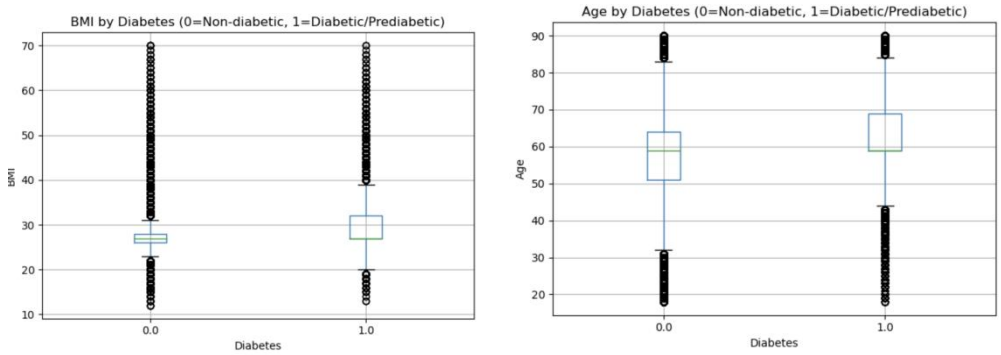
2.4.1 Visualization

Distribution of BMI & Age histograms:

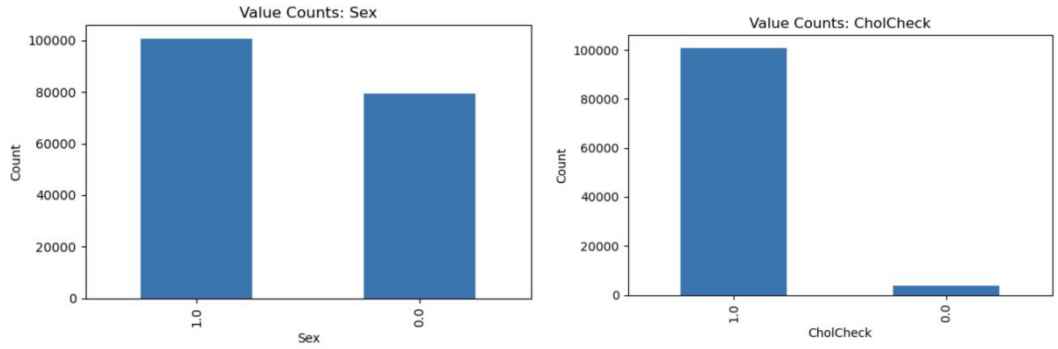




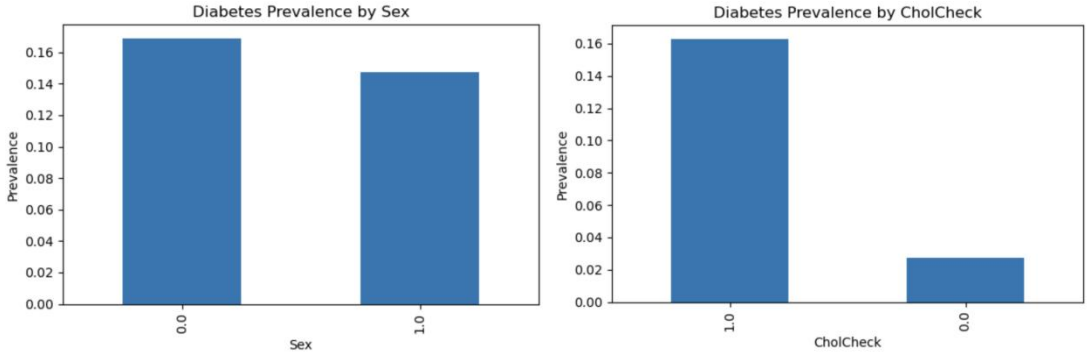
BMI/Age by Diabetes box plots:



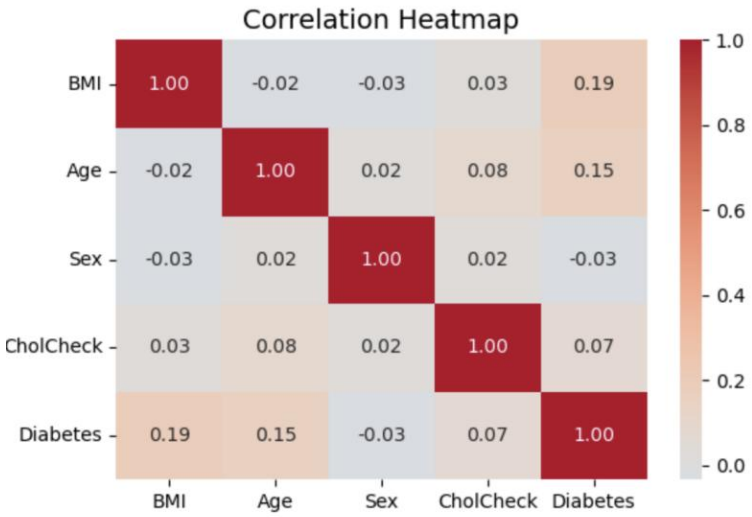
Value counts of Sex/CholCheck bar charts:



Diabetes Prevalence by Sex/CholCheck bar plots:



Correlation heatmap:



2.4.2 Discussion & Practical Implications

BMI: Most participants have BMI between 25–30. Diabetic individuals show higher BMI values, with a correlation of 0.19 to diabetes, making it the strongest predictor among four features.

Age: Older participants have higher diabetes prevalence. The correlation with diabetes is 0.15, confirming age as a key risk factor.

Sex: Men (16.9%) have slightly higher diabetes prevalence than women (14.7%). However, the correlation (-0.03) is weak, suggesting limited impact.

CholCheck: Those who had cholesterol checks show higher prevalence (16.3% vs. 2.7%). This likely reflects that diabetic individuals are more likely to undergo health checks rather than a direct causal effect.

2.4.3 Conclusion

BMI, Age, and CholCheck show meaningful associations with diabetes, while Sex plays a smaller role. To further predict whether an individual is at risk of developing diabetes, we need to take more variables into account.

3. Forest Cover Analysis Report

3.1 Problem Definition

Forest ecosystems play a vital role in environmental conservation. Understanding the distribution of forest cover types is crucial for sustainable forest management, biodiversity conservation, climate change mitigation strategies, and wildlife habitat protection.

3.1.1 Research Question

Do different wilderness areas exhibit distinct patterns of forest cover type distribution? How do topographic features influence these distributions?

3.1.2 Benefits

Forest managers: Optimize resource allocation and conservation planning
Researchers: Gain insight into ecosystem dynamics and species distribution
Policymakers: Data-driven environmental conservation decisions
Conservation organizations: Conduct targeted conservation efforts

3.2 Data Description

The original dataset, after thorough cleaning and validation, contains 30860 samples and 56 features. This dataset represents forest cover types across four wilderness areas in northern Colorado.

	Unnamed: 0	Elevation	Aspect	Slope	Horizontal_Distance_To_Hydrology	Vertical_Distance_To_Hydrology	Horizontal_Distance_To_Roadways
0	0	3351.0	NaN	27.0	726.0	124.0	3813.0
1	1	2732.0	129.0	7.0	212.0	1.0	1082.0
2	2	2572.0	24.0	9.0	201.0	689.0	957.0
3	3	2824.0	69.0	13.0	417.0	39.0	3223.0
4	4	2529.0	84.0	5.0	120.0	9.0	1092.0

3.2.1 Key Characteristic Analysis

Elevation: Range: 1,860 to 3,850 meters
Slope: Terrain steepness (0-90°)
Aspect: Compass direction of slope (0-360°)
Hillshade Index: Sunlight valuesat different times of day
Distance Index: Distance to water, roads, and fires
Wilderness Area Index: Four designated wilderness areas
Forest Cover Type: Seven different vegetation types

```python
#Data Quality Validation Function
def mark_invalid_as_nan(df):
    data_checked = df.copy()
    data_checked.loc[data_checked['Elevation'] <= 0, 'Elevation'] = np.nan
    data_checked.loc[(data_checked['Slope'] < 0) | (data_checked['Slope'] > 90), 'Slope'] = np.nan
    data_checked.loc[(data_checked['Aspect'] < 0) | (data_checked['Aspect'] > 360), 'Aspect'] = np.nan
    hillshade_columns = [col for col in data_checked.columns if col.startswith("Hillshade")]
    for column in hillshade_columns:
        data_checked.loc[(data_checked[column] < 0) | (data_checked[column] > 255), column] = np.nan
    distance_columns = [col for col in data_checked.columns if col.startswith("Distance")]
    for column in distance_columns:
        data_checked.loc[data_checked[column] < 0, column] = np.nan
    return data_checked
```

3.2.2 Data Quality Summary

Initial data cleaning and validation ensured data integrity through the following:
Replacing invalid values with NaN
Data type consistency assurance
Interpolation based on wilderness areas
Comprehensive statistical validation

data quality summary: (55, 6)

	Feature	Dtype	Non_Null	Missing	Missing_Pct	Unique_Count
0	Elevation	float64	25923	4937	16.0	1568
1	Aspect	float64	23763	7097	23.0	363
2	Slope	float64	28083	2777	9.0	58
3	Horizontal_Distance_To_Hydrology	float64	28392	2468	8.0	425
4	Vertical_Distance_To_Hydrology	float64	29009	1851	6.0	505
5	Horizontal_Distance_To_Roadways	float64	28392	2468	8.0	4548
6	Hillshade_9am	float64	21602	9258	30.0	179
7	Hillshade_Noon	float64	27157	3703	12.0	141
8	Hillshade_3pm	float64	22220	8640	28.0	245
9	Horizontal_Distance_To_Fire_Points	float64	24997	5863	19.0	3911
10	Soil_Type1	float64	24997	5863	19.0	2
11	Soil_Type2	float64	28083	2777	9.0	2
12	Soil_Type3	float64	25306	5554	18.0	2
13	Soil_Type4	float64	24071	6789	22.0	2
14	Soil_Type5	float64	27774	3086	10.0	2
15	Soil_Type6	float64	22220	8640	28.0	2
16	Soil_Type7	int64	30860	0	0.0	2
17	Soil_Type8	int64	30860	0	0.0	2
18	Soil_Type9	int64	30860	0	0.0	2
19	Soil_Type10	int64	30860	0	0.0	2
20	Soil_Type11	int64	30860	0	0.0	2
21	Soil_Type12	float64	15122	15738	51.0	2
22	Soil_Type13	float64	23145	7715	25.0	2
23	Soil_Type14	float64	23145	7715	25.0	2
24	Soil_Type15	float64	24380	6480	21.0	1
25	Soil_Type16	float64	24688	6172	20.0	2
26	Soil_Type17	float64	17899	12961	42.0	2
27	Soil_Type18	float64	24688	6172	20.0	2
28	Soil_Type19	float64	14196	16664	54.0	2
29	Soil_Type20	float64	12344	18516	60.0	2
30	Soil_Type21	float64	17899	12961	42.0	2

	Elevation	Aspect	Slope	Horizontal_Distance_To_Hydrology	Vertical_Distance_To_Hydrology	Horizontal_Distance_To_Roadways	Hillshade_9am	Hillshade
0	3351.0	NaN	27.0	726.0	124.0	3813.0	192.0	
1	2732.0	129.0	7.0	212.0	1.0	1082.0	231.0	
2	2572.0	24.0	9.0	201.0	689.0	957.0	216.0	
3	2824.0	69.0	13.0	417.0	39.0	3223.0	NaN	
4	2529.0	84.0	5.0	120.0	9.0	1092.0	NaN	
5	2050.0	284.0	42.0	NaN	141.0	192.0	75.0	
6	3004.0	236.0	5.0	960.0	95.0	5814.0	211.0	
7	3232.0	111.0	10.0	NaN	78.0	1342.0	NaN	
8	3141.0	156.0	6.0	503.0	72.0	NaN	228.0	
9	NaN	NaN	NaN	30.0	3.0	630.0	NaN	

10 rows × 55 columns

3.3 Data Cleaning and Processing

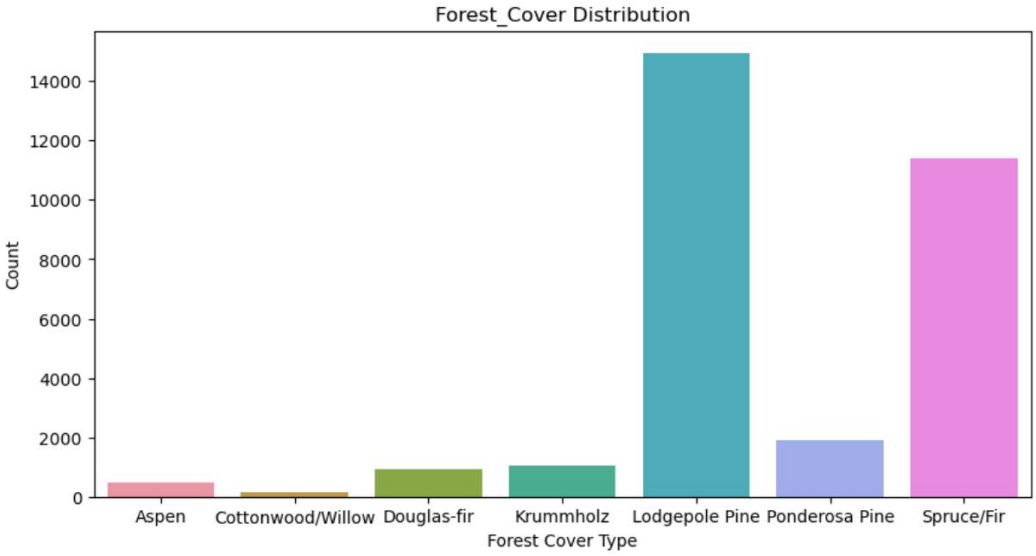
Strict validation rules were applied:
Elevation: Negative values were converted to NaN
Slope: Values outside the range [0, 90] were corrected
Aspect: Values outside the range [0, 360] were processed
Hillshade Index: Values outside the range [0, 255] were processed Out-of-range values
Distance metric: Negative values are handled appropriately

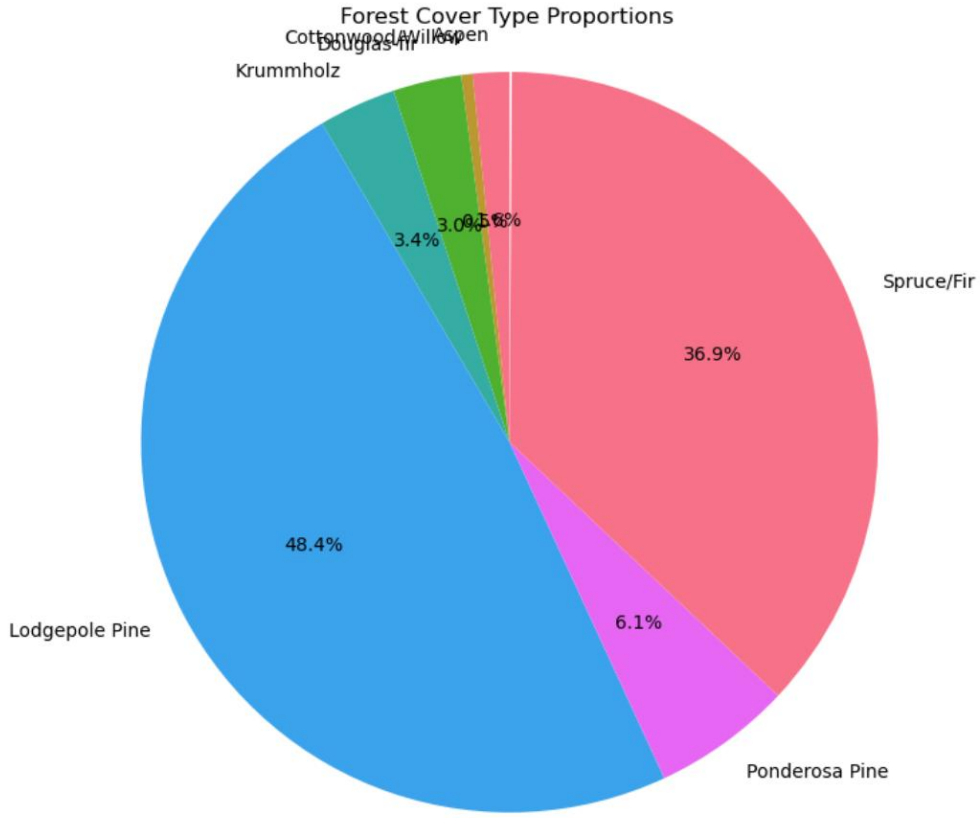
```
Invalid values replaced with NaN (top 10 columns):
Slope              1421
Hillshade_Noon     1350
Aspect             1208
Hillshade_3pm      1143
Hillshade_9am      1081
Elevation           667
Soil_Type24           0
Soil_Type23           0
Soil_Type33           0
Soil_Type25           0
dtype: int64

Dataset shape after validation: (30860, 55)

Sample of validated data (first 10 rows):
```

3.4 Exploratory Data Analysis

3.4.1 Distribution of Forest Cover Types

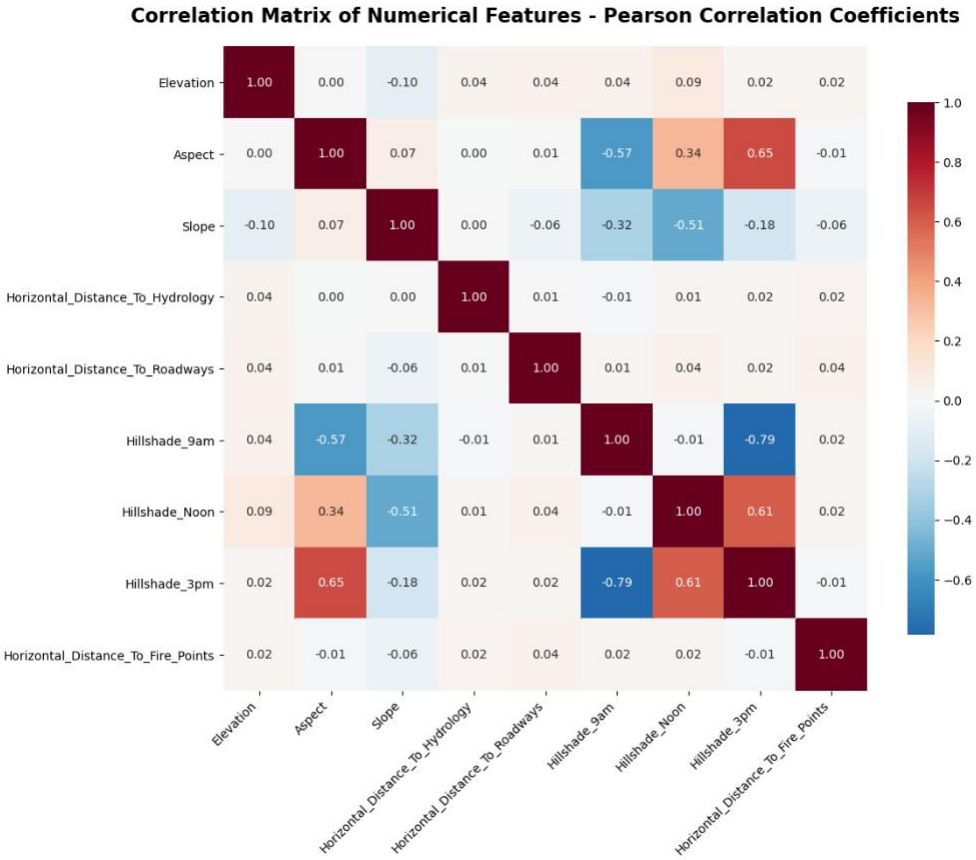
The seven forest cover types exhibited distinct distribution patterns within the study area, with certain types predominant in specific elevation ranges and wilderness areas.





3.4.2 Statistical Analysis of Topographic Characteristics

Comprehensive linear correlation network among 9 key terrain and environmental: "Elevation","Aspect","Slope","Horizontal_Distance_To_Hydrology","Horizontal_Distance_To_Roadways","Hillshade_9am","Hillshade_Noon","Hillshade_3pm","Horizontal_Distance_To_Fire_Points"

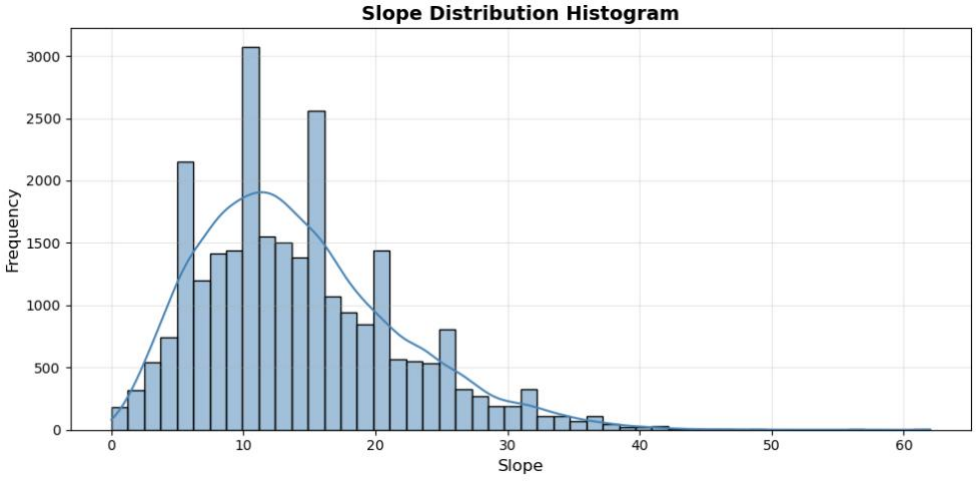


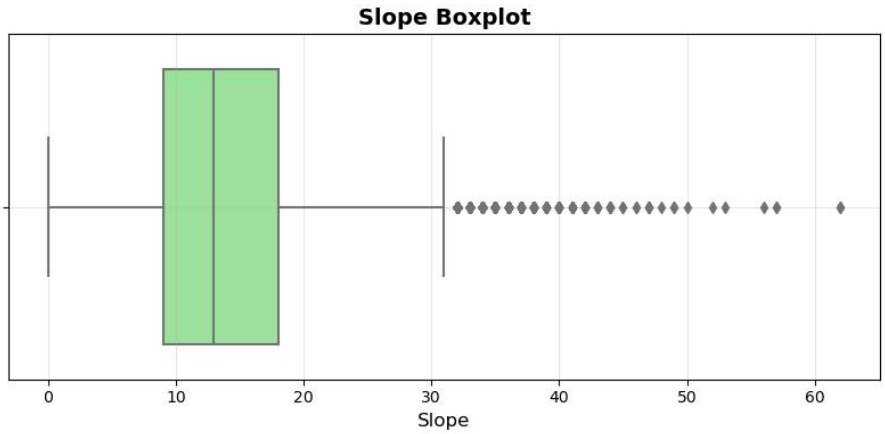
features

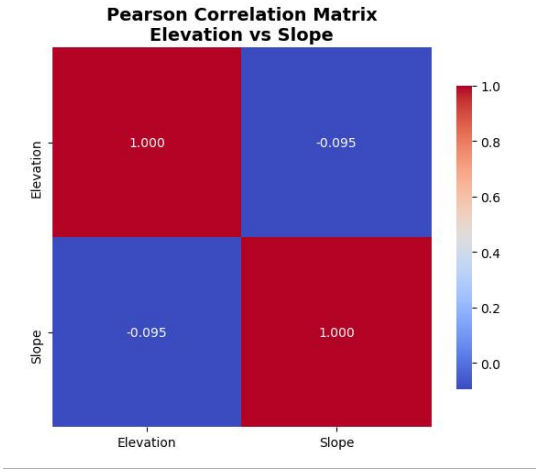
3.4.3 Characteristic Distribution Analysis

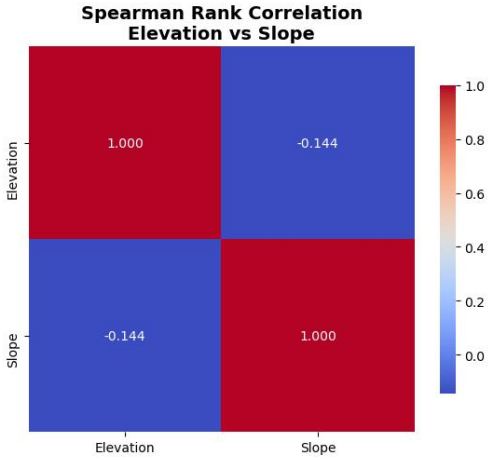
Slope distribution is right-skewed, with most areas having gentle slopes (0-20 degrees).

Pearson linear correlation between the two features Elevation and Slope.









3.4.4 Discussion and Practical Implications

The analysis revealed clear relationships between topographic characteristics and forest cover. High-elevation areas supported distinct vegetation communities compared to low-lying valleys, while aspect and steepness created microhabitats that influenced species composition.

Practical Applications:

Conservation Planning: Prioritize areas with unique vegetation assemblages

Forest Management: Develop specific measures based on the requirements of specific cover types

Climate Adaptation: Identify vulnerable vegetation communities

Research Focus: Target areas of high ecological diversity

3.4.5 Conclusions

This analysis demonstrates a significant association between topographic characteristics and forest cover distribution in wilderness areas. The results support the importance of environmental factors in shaping vegetation patterns and provide valuable insights for sustainable forest management and conservation efforts. Comprehensive data cleaning and analysis methods ensure reliable results that can inform scientific research and practical land management decisions.

References

Carlier, M. (2018). *U.S. new and used car sales 2018 | Statista*. Statista; Statista. https://www.statista.com/statistics/183713/value-of-us-passenger-cas-sales-and-leases-since-1990/

Ebru Caglayan Akay, Ömer Faruk Bolukbasi, & Engin Bekar. (2018). Robust and Resistant Estimations of Hedonic Prices for Second Hand Cars: an Application to the Istanbul Car Market. *DOAJ (DOAJ: Directory of Open Access Journals)*.

Kihm, A., & Vance, C. (2016). The determinants of equity transmission between the new and used car markets: a hedonic analysis. *Journal of the Operational Research Society*, *67*(10), 1250–1258. https://doi.org/10.1057/jors.2016.16